

U. S. ARMY RESEARCH OFFICE

Report No. 86-1

February 1986

TRANSACTIONS OF THE THIRD ARMY CONFERENCE
ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

Host

Georgia Institute of Technology

Atlanta, Georgia

13-16 May 1986

Approved for public release; distribution unlimited.
The findings in this report are not to be construed
as an official Department of the Army position un-
less so designated by other authorized documents.

U.S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

FOREWORD

The Third Army Conference on Applied Mathematics and Computing was held at the Georgia Institute of Technology, Atlanta, Georgia on 13-16 May 1985. The sponsor of these annual meetings is the Army Mathematics Steering Committee (AMSC). Its members would like to thank Professor William F. Aimes for his invitation to hold the conference at his university, and for his outstanding work as the Chairperson on Local Arrangements. He, along with Professors M. F. Barnsley, Albert Bharucha-Reid, and R. W. Schafer organized the three special Sessions for this meeting. Unfortunately Professor Bharucha-Reid passed away before this conference was held, but his help in selecting the speakers and other planning phases contributed a great deal to its success. He was a skillful mathematician whose advice and council was often sought, and to many in the scientific community he was a friend and so as time gives on he will be missed more and more.

The program of the present conference consisted of three parts, namely: (a) Seven invited addresses; (b) Three special sessions; and (c) Contributed papers by Army, academic and other scientific personnel. The sixteen speakers for the special sessions were selected by their organizers. These sessions carried the titles: "Algorithmic Issues in Multi-Dimensional Digital Signal Processing," "Applications of Chaotic Dynamics," and "Invariant Solutions of Partial Differential Equations." The subcommittee of the AMSC that oversee these conferences was very pleased with the high scientific quality of the fifty-one contributed papers. The Army scientists had an opportunity to hear and talk with many nationally known mathematical scientists during the course of this meeting. Some of these were the invited speakers, who are listed below together with the titles of their addresses, but also with many others that appeared on the program or were members of the audience.

SPEAKERS AND AFFILIATION

TITLES OF ADDRESS

Professor J. T. Schwartz
Courant Institute of Mathematical
Sciences

Identification of Partially Obscured
Objects in Two and Three Dimensions by
Matching of Noisy "Characteristic
Curves"

Professor S. Rosenblatt
Illinois Institute of Technology

Bifurcation and Stability of
Viscoelastic Fluid Flows

Dr. V. K. Stokes
General Electric Company

The Role of Modelling in an Industrial
Environment

Professor J. Strickwerda
University of Wisconsin

Finite Difference Methods for
Elliptic Systems

Professor S. N. Atluri
Georgia Institute of Technology

Computational Aspects of Finite Strain
Inelastic Solid and Fracture Mechanics

Professor J. R. Rice
Purdue University

Using Supercomputers: Today and
Tomorrow

Professor M. H. Schultz
Yale University

Parallel Computing and Fluid Dynamics

A large number of individuals contributed to the success of the conference, in particular, the many speakers, the chairpersons, the host personnel, and the active and enthusiastic members of the audience. The members of the AMSC were please that most of the speakers were able to find time to prepare their papers for these Transactions. These research articles will enable many persons that were not able to attend the symposium to profit by these contributions to the scientific literature.

TABLE OF CONTENTS

<u>Title</u>	<u>Page</u>
Foreword	iii
Table of Contents	v
Program	xi
Some Remarks on Robot Vision Jacob T. Schwartz and Micha Shario	1
Category Learning and Adaptive Pattern Recognition: A Neural Network Model Gail A. Carpenter and Stephen Grossberg	37
Nonlinear Neural Dynamics of Visual Segmentation Stephen Grossberg and Ennio Minolla	57
Genie: An Inference Engine with Vulnerability Applications Fred Brundick, John Dumer, Timothy Hanratty, Ralph Shear, and Paul Tanenbaum	75
Analysis of Gradient Change Thresholds in the Detection of Edges of Object from Range Data C. N. Shen and R. L. Racicot	89
Identification of Partially Obscured Objects Charles R. Leake	107
Optimum Control of Flexible Robot Arms on Fixed Paths Sabri Cetinkunt and Wayne J. Book	111
Dynamics of Flexible Mechanical Systems Wan S. Yoo and Edward J. Haug	123
Response of Damped Mechanical Systems to a Time Dependent External Force Gary L. Anderson	147
Computation of Residual Stresses Due to Phase Transformations During Quenching of Hollow Cylinders J. D. Vasilakis	189

*This Table of Contents lists only the papers that are published in this Technical Manual. For a list of all the papers presented at the First Army Conference on Applied Mathematics and Computing, see the Agenda

<u>Title</u>	<u>Page</u>
Further Investigation of the Stability of Diffusion Flames Near Extinction Y. S. Choi and G. S. S. Ludford	213
Complex Kinetics in Flame Theory G. S. S. Ludford and Richard Y. Tam	217
Application of Front Tracking to Combustion, Surface Instabilities and Two Dimensional Riemann Problems: A Conference Report Bruce Bukiet, Carl L. Gardner, James Glimm, John Grove, James Jones, Oliver McBryan, Ralph Menikoff and David H. Sharp	223
Bifurcation and Stability of Viscoelastic Fluid Flows S. Rosenblat	245
Crack Solutions and Ductile Fracture Criteria Dennis M. Tracey and Colin E. Freese	259
The Baushinger Effect on Stress Intensity Factors for a Radial Cracked Gun Tube S. L. Pu and P. C. T. Chen	275
Elastic-Plastic Loading and Unloading in a Thick Tube with Kinematic Hardening Theory Peter C. T. Chen	295
A Simplified Orthotropic Formulation of the Viscoplasticity Theory Based on Overstress M. Sutcu and E. Krempl	307
Linear Stability of Shear Flow of a Viscoelastic Fluid Yuriko Renardy and Michael Renardy	339
Effect of a Wall on the Lift Force Donald A. Drew	345
The Effects of Non-Sphericity and Radiative Energy Loss on the Migration of the Gas Bubble from Underwater Explosions K. C. Heaton	353
Stefan's Problem in a Finite Domain with Constant Boundary and Initial Conditions Shunsuke Takagi	403
Numerical Aberations in a Stefan Problem from Detonation Theory G. S. S. Ludford and A. A. Oyediran	405

<u>Title</u>	<u>Page</u>
The Role of Modeling in an Industrial Environment Vijay K. Stokes	415
Condensation on Fractals Sets J. S. Geronimo	417
Newton's Method, Julia Sets and Chaotic Dynamics Edward R. Viscay	423
Chaotic Eigenstates for Quantum Mechanical Systems D. Bessie	439
Large Deformations of Elastomer Cyclinders Subjected to End Thrust and Probe Penetration A. R. Johnson, C. J. Quigley and I. Fried	449
A Technique for Calculating Path Integrals for Nonlinear Fracture J. R. Whiteman and G. M. Thompson	467
An Explicit and Priori Assessment of Shear Locking in a Triangular Mindlin-Type Plate Element Alexander Tessler	471
On the Solvability and Computational Aspects of a Refined Shear Deformation Plate Theory J. N. Reddy	493
Three Phase Flow in a Porous Medium and the Classification of Non-Strictly Hyperbolic Conservation Laws Michael Shearer and David G. Schaeffer	509
A Generalized Heat Equation: An Overview Siegfried H. Lehnigk	519
On the Numerical Solution of a Stochastic Optimal Correction Problem P. L. Chow and J. L. Menaldi	531
Error Estimation for the Numerical Solution of a Stochastic Control Problem P. L. Chow and J. L. Mendaldi	547
Finite Difference Methods for Elliptic Systems John C. Strikwerda	559
Generalized Isovectors and Similarity Solutions Frank B. Estabrook and Hugo D. Wahlquist	567
Application of Reciprocal Backlund Transformations to Stefan Problems in Nonlinear Heat Conduction Colin Rogers	573

<u>Title</u>	<u>Page</u>
Analysis of Fluid Equations by Group Methods W. F. Ames and M. C. Nucci	589
F-P-S Poincare'-Like Linearization Applied to Soliton Equations R. L. Anderson and E. Taflin	597
Group Analysis of the Pellet Fusion Process V. J. Ervin, W. E. Ames and E. Adams	605
An Endochronic Approach and Other Topics in Small and Finite Deformation Computation Elasto-Plasticity Satya N. Atluri	619
Development of Singularities in Nonlinear Viscoelasticity J. A. Nohel and M. Renardy	639
A Fast Algorithm for Non-Newtonian Flow David S. Malkus	651
Wave Curves for the Riemann Problem of Plane Waves in Simple Isotropic Elastic Solids Zhying Tang and T.C.T. Ting	661
A Comparison Between Vector and Tensor Transformations, An Application in Continuum Mechanics M. N. L. Narasimhan and Edward A. Saibel	669
Large Elastic Deformation of a Sheet Due to Fluid Load Edward W. Ross, Jr.	693
Relativistic Wave Equations for Solids and Low Temperature Quantum Systems Richard A. Weiss	717
Regularity Results for the Porous Medium Equation Klaus Hollig and Heinz-Otto Kreiss	741
On the Treatment of Poisson's Equation by Piecewise Polynomials and Partition Method Shih C. Chu	749
Adaptive, Self-Validating Numerical Quadrature George F. Corliss and L. B. Rall	757
Aspects of a High Level Algorithm for Processing Diverging and Converging Branch Nonserial Dynamic Programming Systems Augustine O. Esogbue and Nazir A. Warsi	783
A Model for Symmetric Vortex-Merger M. V. Melander, N. J. Zabusky and J. C. McWilliams	801

<u>Title</u>	<u>Page</u>
Stable Summation Methods for Elliptic Eigenfunction Expansions	
Harvey Diamond, Mark Kon and Louise Raphael	819
On the Use of Piecewise-Polynomials for the Approximation of Cauchy Singular Integrals	
Apostolos Gerasoulis	825
Numerical Solution of Random Linear Volterra Integral Equations	
M. Sambandham	841
Comments on Finite Element Method and Band-Width Reduction with Reference to Transient Heat Conduction	
R. Yalamanchili	857

THIRD ARMY CONFERENCE
ON
APPLIED MATHEMATICS AND COMPUTING
MAY 13-16, 1985
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332

AGENDA

MONDAY, May 13, 1985

0800-1600 REGISTRATION

0830-0845 OPENING REMARKS

0845-0945 GENERAL SESSION I

CHAIRPERSON: Edward W. Ross, Jr., US Army Natick Research and
Development Laboratories, Natick, Massachusetts

- IDENTIFICATION OF PARTIALLY OBSCURED OBJECTS IN TWO AND THREE
DIMENSIONS BY MATCHING OF NOISY "CHARACTERISTIC CURVES"

J. T. Schwartz, Courant Institute of Mathematical Sciences,
New York, New York

0945-1015 BREAK

1015-1215 SPECIAL SESSION I - Algorithmic Issues in Multi-Dimensional
Digital Signal Processing

CHAIRPERSON: R. W. Schafer, Georgia Institute of Technology,
Atlanta, Georgia

- GAUSS AND THE HISTORY OF THE FAST FOURIER TRANSFORM

D. H. Johnson, Rice University, Houston, Texas

- MATHEMATICAL MORPHOLOGY AND IMAGE PROCESSING

P. A. Maragos and R. W. Schafer, Georgia Institute of Technology,
Atlanta, Georgia

- PROBLEMS IN CONSTRAINED SIGNAL ESTIMATION

R. M. Mersereau, Georgia Institute of Technology, Atlanta, Georgia

MONDAY, May 13, 1985

● RECENT RESULTS ON SIGNAL PROCESSING ALGORITHMS

C. S. Burrus, Rice University, Houston, Texas

1015-1215

TECHNICAL SESSION 1

CHAIRPERSON: Ronald L. Racicot, Benet Weapons Laboratory,
Watervliet, New York

● NEURAL DYNAMICS OF ADAPTIVE PATTERN RECOGNITION: AUTOMATIC
MATCHING, SEARCH, AND CATEGORY FORMATION

Stephen Grossberg and Gail Carpenter, Boston University,
Boston, Massachusetts

● NEURAL DYNAMICS OF FORM PERCEPTION: BOUNDARY COMPLETION AND
PERCEPTUAL GROUPING

Stephen Grossberg and Ennio Mingolla, Boston University,
Boston, Massachusetts

● LEAST SQUARES MODEL FITTING TO FUZZY VECTOR DATA

Aivars Celmins, US Army Ballistic Research Laboratory,
Aberdeen Proving Ground, Maryland

● GENIE: AN INFERENCE ENGINE WITH APPLICATIONS TO VULNERABILITY
ANALYSIS

F. Brundick, J. Dumer, T. Hanratty and P. Tanenbaum, US Army
Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

1215-1345

LUNCH

1345-1545

TECHNICAL SESSION 2

CHAIRPERSON: Norman Coleman, Jr., US Army Armament Research and
Development Command, Dover, New Jersey

● IMPLICITIZATION OF PARAMETRICALLY DEFINED SURFACES

B. Noble, Mathematics Research Center, Madison, Wisconsin and
M. A. Hussain, General Electric Corporate Research and Development
Center, Schenectady, New York

● ANALYSIS OF GRADIENT CHANGE THRESHOLDS IN THE DETECTION OF EDGES OF
OBJECTS FROM RANGE DATA

C. N. Shen and R. L. Racicot, Benet Weapons Laboratory,
Watervliet, New York

MONDAY, May 13, 1985

● MODEL OF IDENTIFICATION OF PARTIALLY OBSCURED OBJECTS

C. R. Leake, US Army Concepts Analysis Agency, Bethesda, Maryland

● OPTIMUM CONTROL OF FLEXIBLE ROBOT ARMS ON FIXED PATHS

S. Cetinkunt and W. Book, Georgia Institute of Technology,
Atlanta, Georgia

● DYNAMICS OF ARTICULATED STRUCTURES

W. Yoo and E. Haug, University of Iowa, Iowa City, Iowa

● RESPONSE OF DAMPED MECHANICAL SYSTEMS TO A TIME DEPENDENT
EXTERNAL FORCE

G. Anderson, US Army Research Office, Research Triangle Park,
North Carolina

1345-1545

TECHNICAL SESSION 3

CHAIRPERSON: Arthur Wouk, US Army Research Office, Research
Triangle Park, North Carolina

● ANALYSIS OF HEAT TRANSFER THROUGH A SURFACE

J. F. Polk, US Army Ballistic Research Laboratory, Aberdeen
Proving Ground, Maryland

● COMPUTATION OF RESIDUAL STRESSES DUE TO PHASE TRANSFORMATIONS
DURING QUENCHING OF HOLLOW CYLINDERS

J. Vasilakis, US Army Armament Research and Development Center,
Watervliet, New York

● FURTHER NUMERICAL RESULTS FOR NEAR-EXTINCTION DIFFUSION FLAMES

Y. S. Choi, C. Laine-Schmidt and G. S. S. Ludford, Cornell
University, Ithaca, New York

● COMPLEX KINETICS IN FLAME THEORY

G. S. S. Ludford and R. Tam, Cornell University, Ithaca, New York

● DETONATION WAVES, TWO DIMENSIONAL RIEMANN PROBLEMS AND INTERFACE
INSTABILITIES

B. Bukiet, C. Gardner, J. Glimm, J. Grove, J. Jones and O. McBryan
NYU Courant Institute of Mathematical Sciences, New York, New York
and R. Menikoff and D. Sharp, Los Alamos National Laboratories

MONDAY, May 13, 1985

- DIFFERENCE SCHEMES IN FRONT TRACKING FOR HYPERBOLIC SYSTEMS
B. J. Plohr, Mathematics Research Center, Madison, Wisconsin

1545-1605 BREAK

1605-1705 GENERAL SESSION II

CHAIRPERSON: Siegfried Lehnigk, US Army Missile Command,
Redstone Arsenal, Alabama

- BIFURCATION AND STABILITY OF VISCOELASTIC FLUID FLOWS
S. Rosenblatt, Illinois Institute of Technology, Chicago, Illinois

TUESDAY, May 14, 1985

0800-1600 REGISTRATION

0830-1030 TECHNICAL SESSION 4

CHAIRPERSON: William Drysdale, Ballistic Research Laboratory,
Aberdeen Proving Ground, Maryland

- CRACK SOLUTIONS AND DUCTILE FRACTURE CRITERIA
D. Tracey and C. Freese, US Army Materials and Mechanics Research
Center, Watertown, Massachusetts
- DUCTILE FRACTURE OF COMPOSITE MATERIALS
R. Barsoum and C. Quigley, US Army Materials and Mechanics Research
Center, Watertown, Massachusetts
- THE BAUSCHINGER EFFECT ON STRESS INTENSITY FACTORS FOR A RADially
CRACKED GUN TUBE
S. L. Pu and P. C. T. Chen, Benet Weapons Laboratory, Watervliet,
New York
- ELASTIC-PLASTIC LOADING AND UNLOADING IN A THICK TUBE WITH
KINEMATIC HARDENING
P. Chen, Benet Weapons Laboratory, Watervliet, New York
- VISCOPLASTICITY THEORY BASED ON OVERSTRESS FOR ANISOTROPIC
MATERIALS
M. Sutcu and E. Krempl, Rensselaer Polytechnic Institute, Troy,
New York

TUESDAY, May 14, 1985

● A PROBABILISTIC VIEW OF TWO MODELS FOR FATIGUE CRACK GROWTH

A. Goss and N. Singpurwalla, George Washington University,
Washington, DC

0830-1030

TECHNICAL SESSION 5

CHAIRPERSON: Miles Miller, Chemical Research and Development
Center, Aberdeen Proving Ground, Maryland

● LINEAR STABILITY OF SHEAR FLOW OF A VISCOELASTIC FLUID

Y. Renardy and M. Renardy, Mathematics Research Center,
Madison, Wisconsin

● STATE SELECTION FOR TAYLOR-VORTEX FLOW

T. Herbert and Ri Lua Li, Virginia Polytechnic Institute and
State University, Blacksburg, Virginia

● MOTION OF A SPHERE NEAR A WALL

D. Drew, Rensselaer Polytechnic Institute, Troy, New York

● MIGRATION OF THE GAS GLOBE FROM UNDERWATER EXPLOSIONS II: THE
EFFECTS OF NON-SPHERICITY AND ENERGY LOSS

K. Heaton, Defence Research Establishment Valcartier, Courcellette,
P. Q.

● STEFAN'S PROBLEM IN A FINITE DOMAIN

S. Takagi, US Army Cold Regions Research Engineering Laboratory,
Hanover, New Hampshire

● NUMERICAL ABERRATIONS IN A STEFAN PROBLEM

A. Oyediran and G. S. S. Ludford, Cornell University, Ithaca, New York

1030-1100

BREAK

1100-1200

GENERAL SESSION III

CHAIRPERSON: James Thompson, US Army Tank-Automotive Command,
Warren, Michigan

● THE ROLE OF MODELLING IN AN INDUSTRIAL ENVIRONMENT

V. K. Stokes, General Electric Company, Schenectady, New York

TUESDAY, May 14, 1985

1200-1330 LUNCH

1330-1730 SPECIAL SESSION II - APPLICATIONS OF CHAOTIC DYNAMICS

CHAIRPERSON: M. F. Barnsley, Georgia Institute of Technology,
Atlanta, Georgia

- APPLICATION TO CHAOTIC DYNAMICS TO COMPUTER OBJECT RECOGNITION
M. F. Barnsley, Georgia Institute of Technology, Atlanta, Georgia
- DYNAMICAL SYSTEMS APPROACH TO ALGORITHMS FOR FINDING EIGENVECTORS
S. Batterson, Emory University, Atlanta, Georgia
- CONDENSATION ON FRACTAL SETS
J. Geronimo, Georgia Institute of Technology, Atlanta, Georgia
- CHAOTIC DYNAMICS AND NEWTON'S METHODS
E. Vrscay, Georgia Institute of Technology, Atlanta, Georgia
- CHAOS AND NONLINEAR VIBRATIONS OF BUCKLED BEAMS
N. Abhyankar and S. Hanagud, School of Aerospace Engineering,
Atlanta, Georgia
- CHAOTIC EIGENSTATES FOR QUANTUM MECHANICAL SYSTEMS
D. Bessis, SACLAY, France and School of Mathematics, Georgia Institute
of Technology, Atlanta, Georgia
- INSTABILITY AND CHAOTIC BEHAVIOUR IN A FREE-SURFACE FLOW
W. G. Pritchard, Mathematics Research Center, Madison, Wisconsin

WEDNESDAY, May 15, 1985

0800-1600 REGISTRATION

0830-1030 TECHNICAL SESSION 6

CHAIRPERSON: Raman P. Srivastav, US Army Research Office, Research
Triangle Park, North Carolina

- INTERACTION OF ROTATING BANDS AND RIFLING GROOVES
H. P. Chen and S. Hanagud, Georgia Institute of Technology, Atlanta,
Georgia and T. Tsui, US Army Materials and Mechanics Research
Center, Watertown, Massachusetts

WEDNESDAY, May 15, 1985

● ON HEATING AND COOLING OF BARRELS

R. Yalamanchili, US Army Research and Development Center, Dover,
New Jersey

● LARGE DEFORMATIONS OF ELASTOMER CYLINDERS SUBJECTED TO END THRUST
AND PROBE PENETRATION

A. Johnson and C. Quigley, US Army Materials and Mechanics
Research Center, Watertown, Massachusetts and I. Fried, Boston
University, Boston, Massachusetts

● USE AND ANALYSIS OF FINITE ELEMENT METHODS FOR PROBLEMS CONTAINING
SINGULARITIES WITH APPLICATIONS TO FRACTURE MECHANICS

J. Whiteman, Brunel University, England

● AN EXPLICIT A PRIORI ASSESSMENT OF SHEAR LOCKING IN A TRIANGULAR
MINDLIN-TYPE PLATE ELEMENT

A. Tessler, US Army Materials and Mechanics Research Center,
Watertown, Massachusetts

● ON THE SOLVABILITY AND COMPUTATIONAL ASPECTS OF A HIGHER-ORDER
SHEAR DEFORMATION THEORY OF PLATES

J. Reddy, Virginia Polytechnic Institute and State University,
Blacksburg, Virginia

0830-1030

TECHNICAL SESSION 7

CHAIRPERSON: Yoshisuke Nakano, US Army Cold Regions Research
and Engineering Laboratory, Hanover, New Hampshire

● THREE-PHASE FLOW IN A POROUS MEDIUM AND THE CLASSIFICATION OF
NON-STRICTLY HYPERBOLIC CONSERVATION LAWS

M. Shearer, Duke University, Durham, North Carolina

● A GENERALIZED HEAT EQUATION: AN OVERVIEW

S. Lehnigk, US Army Missile Laboratory, Redstone Arsenal, Alabama

● ON THE NUMERICAL SOLUTION OF AN OPTIMAL CORRECTION
PROBLEM

P. Chow and J. Menaldi, Wayne State University, Detroit, Michigan

● PROBABILISTIC ERROR ESTIMATES FOR SOME SECOND-ORDER FINITE-
DIFFERENCE SCHEMES

P. Chow and J. Menaldi, Wayne State University, Detroit, Michigan

WEDNESDAY, May 15, 1985

- COMPARISON THEOREMS AND ERROR ESTIMATES

G. Ladde, The University of Texas at Arlington, Arlington, Texas

- WAVE PROPAGATION AND SCATTERING IN RANDOM MEDIA BY METHOD OF DISCONTINUOUS STOCHASTIC FIELD

V. K. Varadan and V. V. Varadan, Pennsylvania State University, University Park, Pennsylvania

1030-1050 BREAK

1050-1150 GENERAL SESSION IV

CHAIRPERSON: Raymond Sedney, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

- FINITE DIFFERENCE METHODS FOR ELLIPTIC SYSTEMS

J. Strikwerda, The Mathematics Research Center, Madison, Wisconsin

1150-1315 LUNCH

1315-1600 SPECIAL SESSION III - INVARIANT SOLUTIONS OF PARTIAL DIFFERENTIAL EQUATIONS

CHAIRPERSON: W. F. Ames, Georgia Institute of Technology, Atlanta, Georgia

- GENERALIZED ISOVECTORS AND SIMILARITY SOLUTIONS

F. Estabrook and H. Wahlquist, Jet Propulsion Lab, Caltech, Pasadena, California

- APPLICATION OF RECIPROCAL BACKLUND TRANSFORMATIONS TO STEFAN PROBLEMS IN NONLINEAR HEAT CONDUCTION

C. Rogers, Georgia Institute of Technology, Atlanta, Georgia

- GROUP ANALYSIS OF THE NAVIER-STOKES EQUATIONS

W. F. Ames, Georgia Institute of Technology, Atlanta, Georgia

- POINCARÉ LINEARIZATION OF SOLITON EQUATIONS

R. Anderson, University of Georgia, Athens, Georgia

WEDNESDAY, May 15, 1985

● GROUP ANALYSIS AND PELLET FUSION

V. Ervin, Georgia Institute of Technology, Atlanta, Georgia

1600-1620

BREAK

1620-1720

GENERAL SESSION V

CHAIRPERSON: Dennis Tracey, US Army Materials and Mechanics
Research Center, Watertown, MA 02172

● COMPUTATIONAL ASPECTS OF FINITE STRAIN INELASTIC SOLID AND
FRACTURE MECHANICS

S. N. Atluri, Georgia Institute of Technology, Atlanta, Georgia

THURSDAY, May 16, 1985

0800-1200

REGISTRATION

0830-1030

TECHNICAL SESSION 8

CHAIRPERSON: San Li Pu, Benet Weapons Laboratory, Watervliet,
New York

● DEVELOPMENT OF SHOCKS IN NONLINEAR VISCOELASTIC MATERIALS

J. Nohel, Mathematics Research Center, Madison, Wisconsin

● A FAST ALGORITHM FOR NON-NEWTONIAN FLOW

D. Malkus, Mathematics Research Center, Madison, Wisconsin

● WAVE CURVES FOR PLANE WAVES IN ISOTROPIC SIMPLE ELASTIC SOLIDS

Z. Tang and T. Ting, University of Illinois at Chicago

● A COMPARISON BETWEEN VECTOR AND TENSOR TRANSFORMATIONS, AN
APPLICATION IN CONTINUUM MECHANICS

M. Narasimhan and E. Saibel, US Army Research Office, Research
Triangle Park, North Carolina

● LARGE ELASTIC DEFORMATION OF A SHEET DUE TO FLUID LOAD

E. Ross, US Army Natick Research and Development Laboratories,
Natick, Massachusetts

● RELATIVISTIC WAVE EQUATIONS FOR SOLIDS AND LOW TEMPERATURE
QUANTUM SYSTEMS

R. Weiss, US Army Engineer Waterways Experiment Station,
Vicksburg, Mississippi

THURSDAY, May 16, 1985

0830-1030

TECHNICAL SESSION 9

CHAIRPERSON: William Jackson, US Army Tank-Automotive Command,
Warren, Michigan

● C^{00} -REGULARITY FOR THE POROUS MEDIUM EQUATION

K. Hollig and H. O. Kreiss, Mathematics Research Center,
Madison, Wisconsin

● ON THE TREATMENT OF POISSON'S EQUATION BY PIECEWISE POLYNOMIALS
AND PARTITION METHOD

S. Chu, US Army Armament Research and Development Center, Dover,
New Jersey

● ADAPTIVE, SELF-VALIDATING NUMERICAL QUADRATURE

G. Corliss and L. Rall, Mathematics Research Center, Madison,
Wisconsin

● A NOTE ON THE RELATIONSHIP BETWEEN THE NYSTROM METHOD AND HYBRID
SOLUTION OF LOVE'S FREDHOLM EQUATION

M. Driscoll and R. Srivastav, State University of New York at
Stony Brook, New York

● SOLUTION BOUNDS IN MATHEMATICAL PROGRAMMING

O. Mangasarian, Mathematics Research Center, Madison, Wisconsin

● A HIGH LEVEL COMPUTING ALGORITHM FOR NONSERIAL DYNAMIC PROGRAMMING

A. Esogbue, Georgia Institute of Technology, Atlanta, Georgia and
N. Warsi, Atlanta University, Atlanta, Georgia

1030-1050

BREAK

1050-1250

GENERAL SESSION VI

CHAIRPERSON: Jagdish Chandra, US Army Research Office, Research
Triangle Park, North Carolina

● USING SUPERCOMPUTERS: TODAY AND TOMORROW

J. R. Rice, Purdue University, West Lafayette, Indiana

● PARALLEL COMPUTING AND FLUID DYNAMICS

M. H. Schultz, Yale University, New Haven, Connecticut

1300

ADJOURN

Some Remarks on Robot Vision¹

Jacob T. Schwartz & Micha Sharir
Courant Institute of Mathematical Sciences
and

School of Mathematical Sciences, Tel Aviv University

ABSTRACT

This paper considers the problem of using 'depth' images of portions of 3-D objects, drawn from a finite vocabulary of potential objects having known geometry, to identify the objects and determine their orientation when the objects are viewed from an unknown angle. Several techniques, including a simple 'probing' method which can be used when at least two object features constrain its orientation, and the use of semi-local invariant parameters of shape are suggested.

A robust boundary-matching method, which determines best partial least-squares matchings rapidly is then described. Finally, techniques for using the silhouette of a polyhedral body to determine its identify and orientation is described.

1. Introduction

The goal of robotics is to develop general-purpose mechanisms having 'operative' intelligence, i.e. that rudimentary level of intelligence which is displayed in the every-day handling of objects in the workplace and the home. For this level of capability to be realized, a robot will need to maintain at least a partial model of its environment internally. Such a model would represent the (known aspects of its) environment in symbolic fashion, as a collection of 'objects' having known shape and orientation. Rigid objects are simplest; however a complete environment model would eventually have to accommodate *flexible* objects like rope, paper and cloth, *liquids*, *soft* objects (e.g. mashed potatoes), *amorphous* objects like dustpiles or heaps of crumbs, etc. Foregoing these interesting but more difficult problems, the following remarks will concentrate on the relatively simple class of rigid objects and on the problem of identifying such objects and determining their orientations so that they can be manipulated by a robot. Of course, manipulation also requires an understanding of object properties and inter-object relationships such as centers of mass, relationships of support, and coefficients of friction, all of which are concepts which a capable robot will have to understand. However, we ignore all these issues to focus on the underlying, still unsolved, problem of how to recognize objects, seen from unknown orientations, and possibly seen as parts of complex multi-object scenes.

¹Work on this paper has been supported in part by Office of Naval Research Grant N00014-82-K-0381, by a grant from the US-Israel Binational Science Foundation, and by grants from the

Object recognition begins with raw perceptions, i.e. pixel arrays. These must be analyzed into objects which, if rigid, can be identified by the shapes of their bounding surfaces (but may also have properties other than their shape which can be used to identify and locate them, including color, albedo, acoustic reflectance, magnetic behavior, visual texture, electrical behavior; indeed, any property that a sensor can detect). To analyze the objects in a robot's environment will be more or less difficult depending on whether the objects which can appear are known *a priori*, and on whether observation can be maintained continuously or only applied occasionally.

- (1) In a highly controlled environment, it may be known that only certain objects, or only objects belonging to known classes, whose members are precisely characterized by a small number of parameters, can be present. If these objects are known to change position only when the robot moves them, and if none of the robot's manipulations miscarry, it may be possible to keep track of the objects, without much sensing, by a kind of 'dead reckoning'. Even if some manipulations miscarry, it may be known (or plausibly assumed) that only certain of the objects will move to unknown positions when an attempted motion fails. It may then be possible to find these objects again by differencing the pre-manipulation scene and its miscarried result, e.g. after dropping coins on a smooth floor one can locate them again by searching visually for shiny raised areas on the floor.
- (2) Even if some of the objects move independently of the robot system, it may be possible to maintain a valid environment model by keeping the moving objects under continuous observation. In that case the robot system may be able to follow the position of all objects at all times, and will, e.g., retain the capability of returning them to their base positions on command. For example, a future home robot system with multiple eyes built into the walls of a house might be able to keep all a family's dishes under continuous observation during a meal or party, after the conclusion of which it could return them to their standard positions (after cleaning).
- (3) When new objects can appear in the robot's environment, they will at first be perceived simply as unexpected surfaces. It will then be necessary to analyze these surfaces into the objects for which they belong. This process may be able to exploit *a priori* knowledge that only objects of certain categories will ordinarily appear in the robot's environment.

To be fully useful, the recognition capabilities described in the preceding pages need to be organized appropriately. What is wanted is basically a pair of procedures. The first of these should be an object acquisition procedure to which a succession of objects can be presented and their identities supplied. The acquisition procedure should record the object shapes in some suitably compressed form and can pre-process these shapes so as to obtain a collection of efficient

discriminating tests for subsequent object identification. The second procedure will then use this data to ingest a scene containing one or more of these objects and convert it into a list of the objects present, each with its identity and orientation.

Going one step further, we can describe the goal of vision-based object recognition system as follows. The system should maintain and continually update a set of 'registers', one for each object observed; each of these registers should at all times contain the position and orientation of the corresponding object. For moving objects, the system should update this information continuously. This implies that after initial scene analysis the system must frequently probe to determine how the objects in the scene have moved, and whether new objects have entered the scene.

The values present in such continuously updated registers can be considered to represent the natural outputs of an advanced robot vision capability since they are just what the system needs to control and manipulate its environment. Practical progress in scene analysis will be defined by the classes of 2-D and 3-D objects for which we are able to make such a symbolic interface to the real world available and reliable.

2. Advantageous Forms of Raw Data

Visual information is most useful if it is given as 3-D visual data, i.e. if true 3-D coordinates are immediately calculated for all points observed. We note that devices which provide such 3-D data, usually based on laser-beam scanning or on use of specially 'structured light' are commercially available already; see e.g. [S79], [Sc83], [T83].

The crucial advantage of 3-D vision is that it allows images to be acquired by arbitrarily many eyes. Whereas to take ordinary (2-D) images acquired by several eyes and combine them is not easy, multiple 3-D images of a single scene combine in a trivial way, since they all refer to surfaces in a common geometric space. This makes it possible to use arbitrarily many eyes, some fixed, others mounted on moving parts of the robot system. (Eyes need to be mounted on the robot itself if either the robot can roam freely, or to ensure that the space near moving portions of the robot is not obscured, either by an intervening object or by parts of the robot itself. This second purpose may require specialized eyes of appropriate form and position.) Note that an 'all seeing' eye system of this sophistication subsumes a quite satisfactory proximity sensor, and makes other forms of proximity sensors superfluous.

Of course, this still leaves open technical questions such as: how to combine separate observations; what to do when surfaces seen by more than one eye differ discernably; and when to reject an interpretation because of unacceptably large discrepancies. Nevertheless 3-D images are basically favorable for combination, while 2-D images are basically much less favorable.

3. Image Interpretation Techniques

The visual data gathered by a 3-D sensor, i.e. '3-D' or 'depth' images, can be grouped in a table listing all points in 3-space which lie on some reflecting surface of one of the objects present in a scene. However, since all sensor-gathered data is partly corrupted by noise, acquisition of 3-D images leads at once to the problem of how to identify objects, given slightly noise-corrupted images (or, in other cases, silhouettes) of them.

Whether depth images or silhouettes are in question, we shall assume that the objects present in the scene to be analyzed are drawn from some known collection of possible objects O_1, \dots, O_n . This assumption makes the image analysis problem 'objective' rather than 'psychological': one just wants the computer looking at a scene to calculate an integer or a finite set of integers that tells us exactly which of a known list of possible candidate objects it sees, and from what angles it sees them. However, our simplifying assumption still leaves us free to consider any one of a scale of image interpretation problems of gradually increasing difficulty, all of which are 'objective' in the sense just mentioned, and all of which would contribute robot capabilities of practical significance if solved:

- (1) The bodies present in the scene can be wholly visible or may be partially obscured.
- (2) The bodies can be straight-sided (polygonal or polyhedral), or curved.
- (3) If the bodies present in the scene are polygonal or polyhedral, they may either be known to lie in some constrained orientation, (e.g. standing on an edge or face, atop a flat surface), or can be present in completely arbitrary orientations.
- (4) We may be able to assume that the scene contains just one object, or may need to deal with scenes containing multiple objects, which may have either a single uniform orientation, or many different orientations.
- (5) Instead of fixed objects, we may need to deal with objects which are known only to belong to one of a finite sequence of object classes $O_1(s), \dots, O_n(s)$, each of which depends on one or more shape parameters s (e.g. in a home robotics application, cylindrical cans of various heights and widths may be encountered).

This list defines a family of problems for whose solution appropriate algorithmic or heuristic approaches are needed. The efficiency of the approach selected will be important, especially in the dynamic case where the system needs to keep track of moving objects in real time.

The remarks which follow will describe various semi-algorithmic heuristics of gradually increasing complexity which can be used to handle some of the problems listed above. Some of these approaches have been simulated, and where possible we will note the results of simple numerical experiments. The approach proposed is related to one explored in a series of papers by Y. Shirai of the Tsukuba

Electrotechnical Laboratory in a series of papers; see [OS75], [OS79], [S79], [SKOI83], [SS71].

4. Recognition of 2-D Objects

Two basic approaches to recognition of 2-D objects drawn from a finite collection of candidate objects can be proposed. The first approach applies successive 'probes' to the object, gathering sufficient information to discriminate it from among other potential objects (and to determine its orientation). This approach deals particularly easily with simple polygonal objects, but sidesteps the issue of shape description. The second approach associates a global shape descriptor with each object viewed, and then matches this descriptor to pre-calculated similar descriptors of the model objects expected to be seen.

Recognition by Probing

By processing raw image data in simple ways we can apply various logical 'probes' to it. (In the absence of visual data, probing can be accomplished mechanically by detecting object contact using touch sensors.) Each such probe can be considered to move along some specified curve γ from a given position in a given direction until the first intersection of γ with an object present in the scene is detected. The curves along which one probes can be straight lines, circles, etc. If only silhouettes of an object are given, we can still think of 'silhouette probes', i.e. probes in the silhouette plane which end on encountering a point of the silhouette boundary.

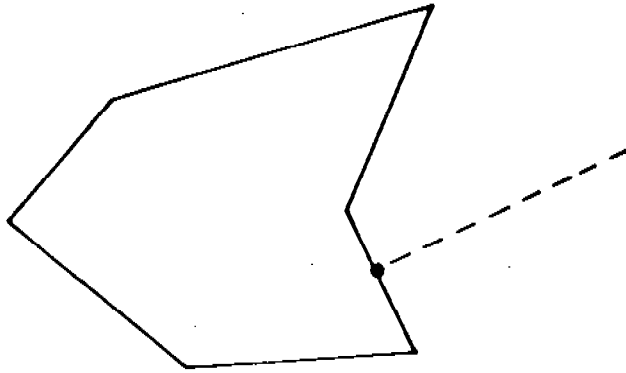


Figure 1(a): Probe of a 2-D polygonal object

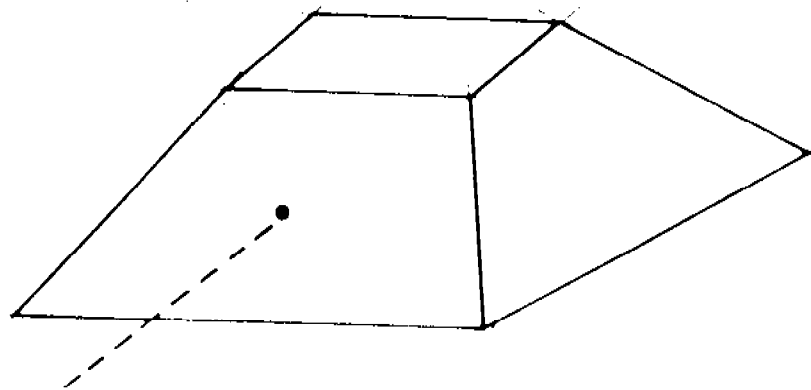


Figure 1(b): Depth probe of a 3-D object

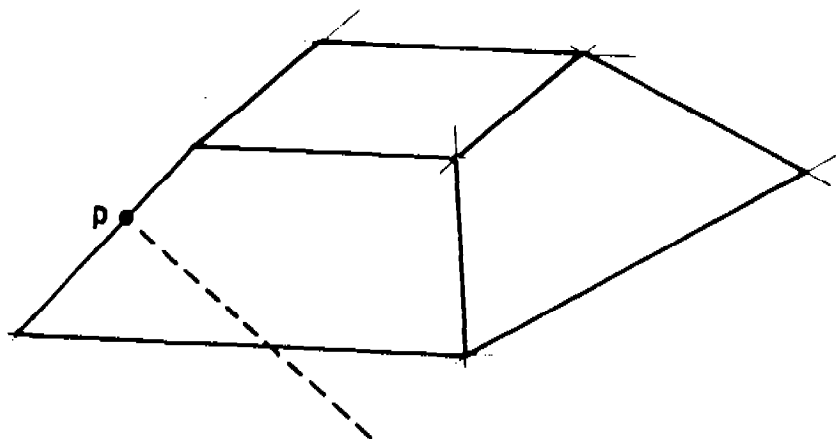


Figure 1(c): Silhouette probe of a 3-D object
(Distance from camera of point p need not be known.)

To see how easily objects can sometimes be identified by probing, consider the simplest 2-D case, in which we are given a polygon standing on one of its edges. In this particularly elementary situation, the first probe conducted will establish a point against which the polygon can be considered to 'fit', and then, knowing the point at which the probe contacts the polygon, we know that the polygon's possible positions are restricted to a finite set. This allows us to build the finite set of points at which a further probe line could intersect one of the polygons which might confront us, in one of its finitely many possible orientations. Suppose we divide this probe line into minimal resolvable intervals (determined by the precision of the instrument with which probes are conducted). Count the number of such intervals which contain points of intersection and calculate the entropy of the associated subdivision; this is the *resolving power* of the probe line. For efficiency one will then want to probe along the line whose resolving power is greatest. Only a subset of the original set of polygons and orientations will remain as candidates after this first probe, and then one can apply a second probe which has greatest resolving power for this subset, etc. The tree-like search which results will determine the identity of the observed polygon and the edge that it is standing on. Normally very few probes will be required. The first probe should be at a level which minimizes the expected number of probes subsequently required. (This style of searching associates a notion of 'entropy', relative to the imprecision of the probing instrument, with the given set O_1, \dots, O_k of objects; this 'entropy' is likely to have interesting invariance properties, and deserves closer study.)

Note that the probing procedure outlined is independent of any assumption of convexity.

Next consider the somewhat less trivial case of a convex polygon whose initial orientation is totally unknown (but assume that we know one point interior to it). Probe the polygon twice to determine two points on its periphery, and then track the segment between them to determine whether these two points belong to the same polygon edge (we assume that such a 'generalized probing' operation is available). If not, probe at a point intermediate between the two first points, and repeat. Eventually we must find two points which lie together on the same edge of the polygon; this reduces us to the case considered previously, since the polygon may be considered to be 'standing' on this edge.

Similar ideas can be applied to the more interesting case of a curved 2-D region. To avoid complications suppose first that the region is convex. If we can locate one point P fixed relative to the region, then, by probing along a circle about this point as center, we can orient the region so that a chord through the point P having a standard length D can be regarded as horizontal. This standardizes the position of the region to one of a finite collection of possible positions, and then we can use the kind of 'probe tree' described above to determine which one of these possible orientations it has.

If the whole of a region is visible, its centroid can serve as such an anchor point. Similar use could also be made of the two most distant points of the region, of the point of the region most distant from the line connecting these two points, etc. Polygon corners can obviously be used as anchor points; acute corners, of which only a few can exist, are obviously preferable to obtuse corners. It is only necessary that any point which probing might identify with a particular anchor point P should belong to some relatively small known set of points fixed relative to the region, but when the anchor point is ambiguous, the probe tree which we build up must reflect all the possible points that might be confused with it.

Next consider the case of partially-obscured 2-D objects. The preceding observations suggest that the first step in recognizing partially obscured objects should be to define noise-immune anchor points which can be located even if part of the region is obscured. This case is of course more difficult than that in which the whole object is visible, because

- (1) For totally visible objects, obvious 'global' anchor points such as the object centroid are available, while for partially occluded objects anchor points must be calculated from relatively 'local' data.
- (2) If the object being observed is nonconvex, then the (boundary of the) convex hull of the portion being observed need not lie on the convex hull of the whole object (see Fig. 11).

Of course, if the observed portion of the object has sharp discriminating features such as acute corners, then finding an anchor point will be relatively easy. If no such sharp features are present, the problem becomes more difficult. One way of approaching it is by associating the object with some appropriate,

geometrically defined function on the unit circle/sphere of directions in 2-D (resp. 3-D) space. The function must be one whose geometric definition makes it invariant with respect to Euclidean motions of the region to be analyzed. Whenever such an artificial 'color' shows sharp transitions or peaks, these can be used to define the anchor points that we need. In effect, this notion of 'artificial color' converts the shape recognition problem into the problem of recognizing 'colored beachballs' when these are seen from an unknown orientation, a problem for which the presence of spots or regions of sharply defined color will clearly be significant.

Artificial colors of the type proposed can be defined in very many ways, but we want to choose one which has peaks or which varies sharply in the vicinity of geometrically significant boundary features of the body to be analyzed. One possible scheme is as follows: take a modified "carpenter's square" MCS, consisting of two half-lines making some standard angle α ($\alpha = 90^\circ$ would be the standard carpenter's square) and fit it over the region so that both of its two sides touch the boundary of the region. The point at which this contact occurs is determined by the orientation θ of (some distinguished one of) the sides of MCS.

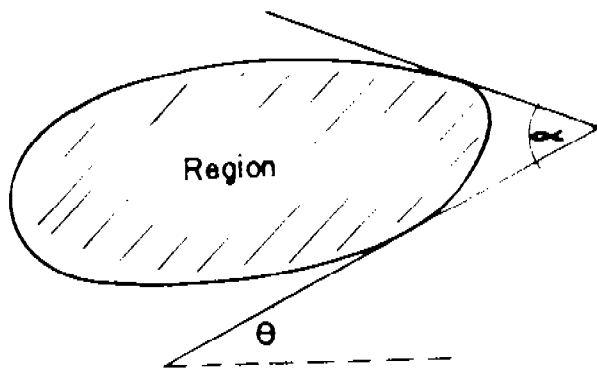


Figure 2: Modified 'carpenter's square' in contact with two points of a body. 'Leading' side is at angle θ to the horizontal.

Let A be the apex of the modified carpenter's square MCS. Take the segment connecting the two points of contact between the region and MCS, take the midpoint M of this segment, and then find and record the distance $d(\theta) = d(\theta, \alpha)$ between the point x at which the line from A to M crosses the region boundary.

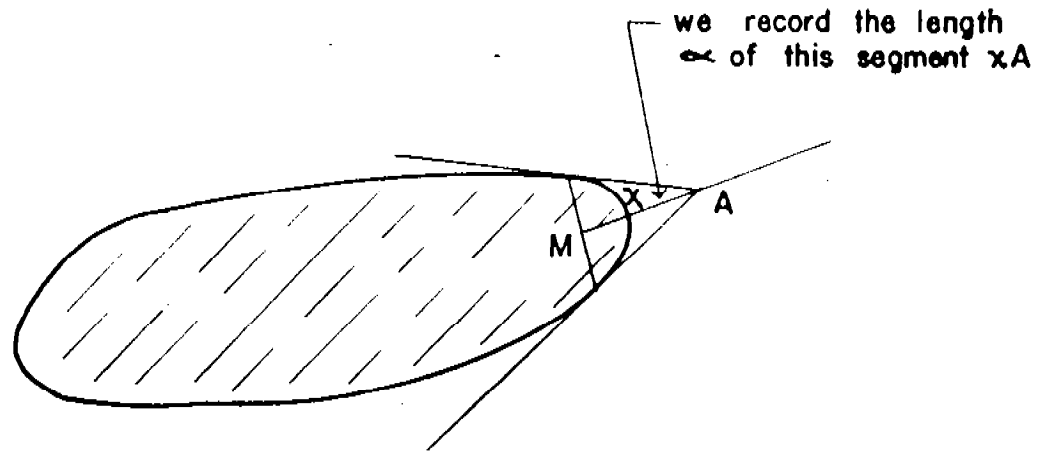


Figure 3: Using a modified carpenter's square to measure a region's average boundary curvature.

If the region were simply a circle of radius R , then the distance $d(\theta) = d(\theta, \alpha)$ would be independent of θ , and would in fact be $R(\sin \alpha/2)^{-1} - 1$. Thus $d(\theta)$ measures a kind of average of the curvature of the periphery of the region; averaged, that is, over the section of periphery between its two points of contact with MCS.

The angle α that can be used in measuring the periphery of a partially obscured region depends on what portion of the periphery is visible. To use an angle α , the tangent to the visible portion of the periphery must turn through an angle exceeding $180^\circ - \alpha$. The closer α approaches 180° , the closer $(\sin \alpha/2)^{-1} - 1$ comes to 0, and hence the more sensitive $d(\theta)$ becomes to small measurement inaccuracies.

If $d(\theta)$ is constant (for several values of the apex angle α of our modified carpenter's square MCS), then the region (or rather the visible portion of its periphery) must be circular, and hence actually possesses no geometric features other than its radius. If $d(\theta)$ is nearly constant, i.e. if the ratio of its largest to its smallest values lies near 1, then the (visible part of the) region will be nearly circular, and hence relatively featureless geometrically. Otherwise this ratio will vary more substantially, enabling us to locate anchor points relatively sharply. To expand upon this remark, it is convenient to consider not $d(\theta)$ but its logarithm $D(\theta) =$

$\log d(\theta)$. By assumption, $D(\theta)$ varies substantially from its minimum value (over the visible part of the periphery, which corresponds to a range of angles $\leq 2\pi$). Suppose that the smallest change in D that we feel able to measure is a change ϵ . Establish a succession of levels $\delta, \delta+\epsilon, \delta+2\epsilon, \dots$ through the range of D . For each of these levels δ_i , divide the range over which θ varies into disjoint intervals, each containing a point at which D takes on the value δ_i , and each terminated by the first occurrence of a sufficiently large interval in which D dips below $\delta_i - \epsilon$ or rises above $\delta_i + \epsilon$. Choose one representative point in each such interval, take this as an anchor point, and make corresponding entries in a probe tree.

To identify a region using this information, we can subsequently survey it with a generalized carpenter's square of appropriate apex angle (depending on the amount of unobscured periphery available, which is appropriately measured in terms of the number of degrees through which the periphery has turned.) Once having measured the boundary in this way, find intervals as above; that is, intervals each of which contains at least one point θ for which $D(\theta) = \delta_i$ and terminated in the same way as the intervals used to build the probe tree. Examine these intervals for each level δ_i , and take the smallest; this gives the most definite information concerning the location of the corresponding anchor point. Then divide this interval of orientations into subintervals, each small enough so that no point of intersection with a probe line can move by more than the standard measurement uncertainty of a probe when the object turns through a single orientation step. In effect, this rule defines the number of 'micro-facets' into which our procedure must divide the interval.

Moving through this range of orientations by stepping successively between the intervals into which we have divided it, take the point x of Fig. 3, which lies between the point M and the apex A of the GCS, as an anchor point, and then execute (or simulate) a series of probes. This will eliminate incorrect orientations/identifications, normally quite rapidly, and leave only those orientations consistent with the available data concerning the visible periphery.

Note that the number of orientations over which we need to search serially will be roughly proportional to

$$\min |I| / \text{var } (D|I)$$

where I designates an angular subinterval of the visible range of tangent angles (to the region periphery), $|I|$ is the size of this subinterval, and $D|I$ designates the restriction of the function D to the subinterval I . Thus favorable cases are those in which a substantial part of the variation of D takes place in some small range of angles; unfavorable cases are those in which D varies uniformly over the whole of the visible angular range. Even in this unfavorable case, the range of angles we have to search will be limited to a fraction of the total angular range inversely proportional to D 's total variation.

The following additional technique can be used to improve the efficiency of the simple approach just outlined. Suppose that the function(s) D using which we are trying to identify and orient a region are constant or nearly constant over some substantial portion B of a region boundary. Then this section B of the boundary is likely to be close to circular, and of a known radius R . We can exploit this fact by mapping the visible portion of the boundary to a much smaller curve. This can be done by moving each of its points P a known distance d (easily calculated from the estimated radius R) perpendicularly away from the tangent line at p . The image of B is then a significantly smaller curve B' . (The image of a perfect circle would plainly be the unique point fixed relative to the circle, i.e. its center.)

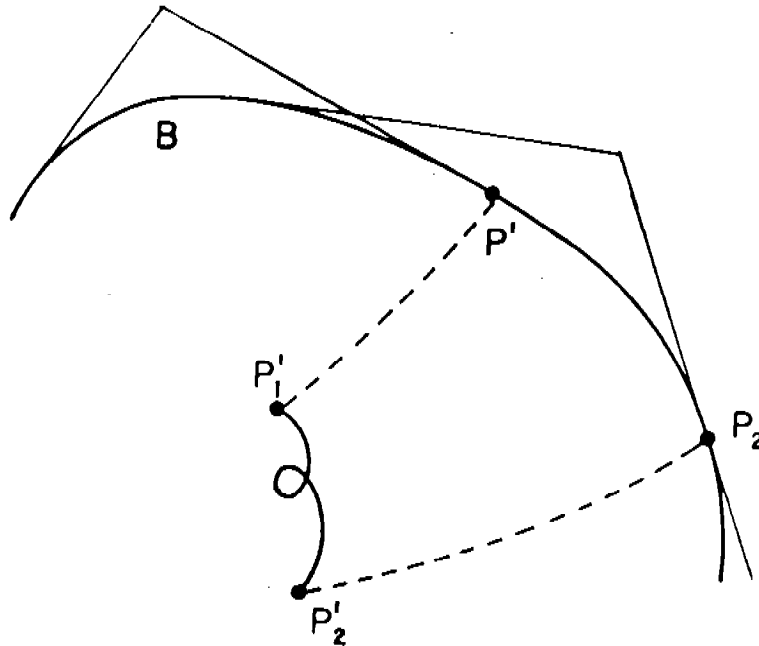


Figure 4: Mapping a nearly circular curve into a smaller curve by 'radial' translation of its points.

Once B has been constructed, we can cover it with sufficiently small circles, which, hopefully, will not be very numerous. The centers of these circles form a collection A of possible anchor points, in the sense that when the curve is measured (with a generalized carpenter's square) and found to have a D -value which only limits the region orientation to the large angular interval B , any point constructed in the manner just explained must lie very close to one of the anchor

points in A.)

5. Some observations concerning geometrically 'colored' and 'colorless' curves and surfaces in 2 and 3 dimensions

The issue crucial to some of the region identification techniques outlined above is how to find one or more 'anchor points' which can be used to standardize the position of the region. (Similarly, flat sides of a region define 'anchor orientations'.) Once such an anchor point has been found, the identification problem becomes very much easier. An anchor point may be unique, or, as in the case of a polygon, many possible points (vertices) may define useful anchors. Moreover, anchor points may be uniquely identified by geometric invariants associated with them, or, as in the case of a regular polygon, a region may possess symmetries and therefore possess multiple anchor points which fall into logically indistinguishable categories.

As we have noted, as soon as a curve is 'painted' with some concrete or abstract 'color' which has significant variation along the curve, it becomes easy to define anchor points; it suffices to take those points having some characteristic color, (but for this we want to pick a color which occurs only infrequently on the curve.) The rotational invariants occurring in the preceding discussion give us a way of operating in situations in which no external color is available, by forming noise-immune geometric invariants and using them as generalized colors. (These 'geometric colors' are most naturally associated with orientations of a measuring instrument and thus can most naturally be regarded as painting the circle rather than the region boundary under investigation. Since, in the case of convex bodies, each orientation maps naturally to a point of the region boundary, this viewpoint loses no significant information.)

We can best understand the potential of this approach by considering those situations in which it must fail. These are situations in which the boundary curve being measured is completely 'colorless' relative to the geometric invariant calculated, i.e. cases in which the battery of invariants we bring to bear have constant values over the boundary of the object being measured. Note that these are also cases other shape matching techniques will also tend to fail, because the same degree of matching will be attained by a large family of orientations differing simply by Euclidean motions, making it impossible to discriminate between these orientations.

To be satisfied with a collection of geometric invariants, we will want constancy of the geometric invariants used to 'paint' an object's boundary to imply that the boundary is *inherently colorless* geometrically, i.e. to imply that its points are equivalent to each other under a Euclidean motion of the whole plane. Curves having this property must clearly be orbits of points under 1-parameter subgroups of the group of plane motions, and hence must either be circles or straight lines (note therefore that if we can see the whole boundary of an object, the circle is the only possible colorless curve). Let us call a set of geometric

invariants *ample* if any curve for which these invariants are constant over the length of a curve is necessarily straight or circular. (In addition, we want invariants that are *stable* relative to small perturbations of a curve, and which are *local*, allowing them to be calculated for nearly the full angular range through which a partially obscured convex curve turns.) Once we have an ample set of invariants (also possessing the other properties just noted) we will have done as well as we can, in the sense that invariants better in any ideal sense are impossible. Similar considerations apply to curves in 3-space and to curves which lie in other geometric objects of concern to us, particularly curves on the sphere.

A similar of geometric 'colorlessness' applies to curves in 3-space and to surfaces. A geometrically colorless curve in 3-space is either a straight line, circle, or helix. A colorless curve lying on the surface of the sphere is necessarily a circle (not necessarily a great circle). A similar notion and remark apply to colorings of the sphere; such a coloring fixes a point (which can then be used as an anchor point) unless there exists a continuous group of rotations of the sphere which leaves the coloring invariant, i.e. $c(Rp) = c(p)$ for the color (or colors) c and every R in some continuous group of rotations. Here there are only two possibilities: either c is constant, or c is constant on each of a family of parallel circles on the sphere. In all other cases, either changes in the shape of one of the level curves $c(p) = \text{const}$ will fix a point, or changes in the relative position of two level curves fix such a point. (For example, for each point p on a first (circular) level curve $c(p) = \text{const}_1$ we can take its minimum distance to a second (also circular) level curve $c(p) = \text{const}_2$; unless the two curves are parallel, this function paints a varying geometric 'color' along the first curve, and (assuming infinite precision) this color fixes an anchor point.

Next consider surfaces in three dimensional space which are geometrically colorless, either in the strong sense that all their points are geometrically equivalent, or in the weaker sense that the surface is invariant under some one-dimensional continuous subgroup of the Euclidean group. In the first case, the surface must have constant principal curvatures, and hence must be a portion either of a plane, sphere, or circular cylinder. In the second case, the orbit of any point under the group of motions leaving the surface invariant must be either a straight line, circle, or helix (of pitch determined by the group leaving the surface invariant). Hence the surface must be either a portion of a cylinder (not necessarily circular), a surface of rotation, or a 'helical cylinder' (screw surface) defined by its cross-section in a plane perpendicular to the direction of the common helix axis.

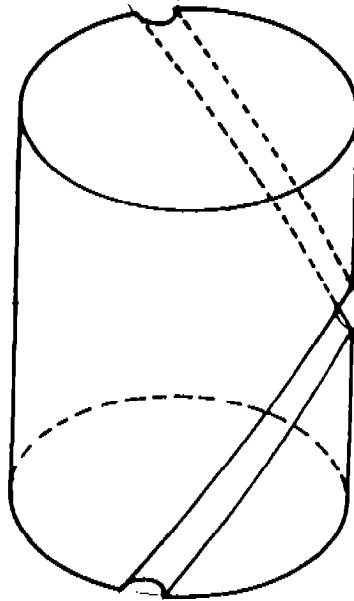


Figure 5: Example of a 'helical cylinder'.
All points of each helix on the surface have equal color.

While interesting as mathematical examples, helical cylinders of this kind are (except for machine screws and their colored equivalent, barber poles) rare.

6. Shape Descriptor Matching

Probing methods like that described above use the actual periphery of the region to be identified, and do not embody any global concept of region shape. This contrasts with other identification techniques that work from some abbreviated shape descriptor which can be associated with the periphery of a convex region, rather than from the periphery itself. The stability and efficiency with which these descriptors can be matched is crucial for such methods. A few mathematical observations can be made concerning this point. Assume first that the observed region is expected to be convex. Such regions are often described by their 'turning function' $\theta(s)$, i.e. the function which, starting from some arbitrarily designated point of its periphery and proceeding counterclockwise around the periphery, records the change in angle of the counterclockwise tangent as a function of the arc-length s traversed. This function is monotone increasing, and varies through 2π as s goes from 0 to its final value S , which is the total periphery

of the region. The value S simply describes the total size of the region, and (if the whole periphery is available) we can normalize it to 2π , so that $\theta(s)$ is monotone and goes from $(0,0)$ to $(2\pi,2\pi)$.

The function $\theta(s)$ has various useful properties:

- (1) $\theta(s)$ is invariant under any Euclidean motion of the object O in question.
- (2) $\theta(s)$ depends in a very simple way on the starting point on the boundary of O , that is, if the starting point shifts by s_0 , the graph of θ undergoes a corresponding horizontal and vertical shift, i.e. simply changes to

$$\theta'(s) = \theta(s + s_0) - \theta(s_0)$$

We can still use the shape descriptor $\theta(s)$, measured along the visible portion of O 's boundary, even if O is partially occluded, i.e. even if only the portion of O which lies right of some (known) directed line is visible. In such case the graph of θ will simply be an (appropriately shifted) portion of the graph for the whole boundary of O .

- (3) θ is parametrized by the arc length of the boundary of O , which can become unstable under small perturbations if convexity is lost. That is, if we represent the noise-corrupted boundary of an observed object O simply as the polygonal line passing through all observed boundary points, the resulting arc length can differ greatly from that of the ideal, noise-free object, in which case the shape descriptors for the observed body and for its model counterpart will not approximate one another. To overcome this difficulty, we can compute the convex hull of the observed data points, obtaining a convexified observed object, and then match the shape descriptor for this convexified observation to pre-stored data describing various model convex bodies.

If the region is polygonal, the function $\theta(s)$ is a step function whose discontinuities tend to be troublesome when a slightly perturbed measurement of $\theta(s)$ is matched against a pre-stored model. To avoid this problem, one can simply turn the graph of the function 45° clockwise, thereby converting it into the graph of a revised function $\eta(s)$ whose derivative is bounded by 1 in absolute value.

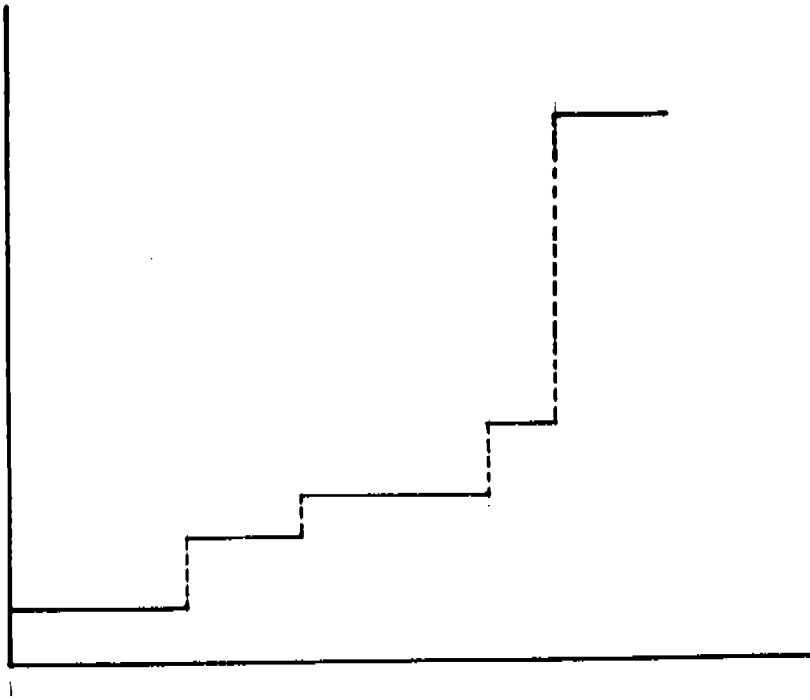


Figure 6(a): Graph of $\theta(s)$ for a polygon.

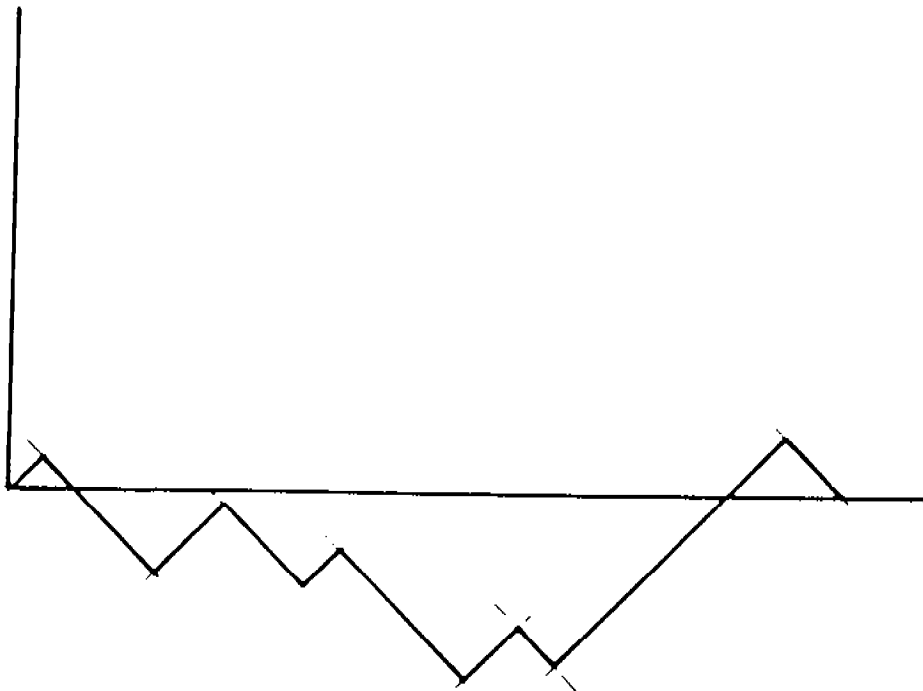


Figure 6(b): Graph of $\theta(s)$ after rotation by 45° .

To identify an observed convex region O (which can be partially obscured), we can then compute its descriptor $\theta(s)$ and try to match it to similar shape descriptors computed for model regions O_1, \dots, O_n corresponding to the various 2-D objects which might appear in an observed scene.

In situations of this kind, matching is customarily implemented by means of the fast Fourier transform algorithm. More specifically, let $\xi(s)$ denote the graph of boundary turning-angle vs. arc length computed for a (convexified) observed polygon O after this graph is turned clockwise by 45° (if O is obscured by a line l then the graph of ξ should start at one point of obscuration (i.e. one intersection of l with the boundary of O) and end at the other such point). For each of the specified model objects O_j let η_j denote the corresponding (rotated) graph for O_j .

If we can use the L^2 metric in shape descriptor space, the object O_j matching O most closely is that which minimizes the distance

$$(2) \quad \min_{d_0} \int_0^L |\eta_j(x+d_0) - \eta_j(d_0) - \xi(x)|^2 dx$$

Actually, we find it better to change (2) slightly so as to find the best "vertical fit" between $\eta_j(x+d_0)$ and $\xi(x)$, i.e. to minimize

$$(3) \quad \min_{d_0} \min_c \int_0^L |\eta_j(x+d_0) - \xi(x) - c|^2 dx$$

In (3) the best value of c is given by

$$(4) \quad c = c(d_0) = \frac{1}{L} \int_0^L (\eta_j(x+d_0) - \xi(x)) dx.$$

With this value of c , (3) becomes

$$(5) \quad \min_{d_0} \left(\int_0^L |\eta_j(x+d_0) - \xi(x)|^2 dx - L|c(d_0)|^2 \right) =$$

$$\min_{d_0} \left(\int_{d_0}^{d_0+L} |\eta_j(x)|^2 dx + \int_0^L |\xi(x)|^2 dx - 2 \int_0^L \eta_j(x+d_0) \xi(x) dx - L|c(d_0)|^2 \right),$$

where in the third integral in the last form of (5) we take ξ to be defined as zero outside the interval $[0, L]$. This allows the minimum appearing in (5) to be rewritten as

$$(6) \quad \min_{d_0} \left(I_j(d_0 + L) - I_j(d_0) - \frac{1}{L} [K_j(d_0 + L) - K_j(d_0) - \int_0^L \xi(x) dx]^2 \right.$$

$$\left. + \int_0^L |\xi(x)|^2 dx - 2 \int_0^L \eta_j(x+d_0) \xi(x) dx \right),$$

where

$$(7) \quad I_j(d) = \int_0^d |\eta_j(x)|^2 dx$$

$$K_j(d) = \int_0^d \eta_j(x) dx.$$

Since the most expensive part of the computation (6) is simply a convolution, it follows that, after discretization to n interpolating points, we can calculate the minimum (6) (for each j separately) in time $O(n \log n)$, using the fast Fourier transform technique.

Direct Match of Rotated 2-D Objects

A cruder but stabler and still quite effective 2-D shape matching scheme can also be implemented efficiently using the fast Fourier transform. In this method, we simply take a sequence of points equally spaced along the perimeter of a (convexified) observed polygon O . More precisely, we take a sequence (u_1, \dots, u_n) of points in clockwise order along the boundary of the convex hull of O such that all the arcs between successive points u_i and u_{i+1} have equal lengths (which must therefore be S/n , where S is the total length of the periphery of O). We then wish to match two such sequences $(u_j)_{j=1}^n$ and $(v_j)_{j=1}^n$ corresponding to an observed (convexified) object O and a model object M respectively. Assume first that the whole boundary of O is visible. Matching amounts to finding a Euclidean motion E of the plane which will minimize the L_2 distance between the sequences $(Eu_j)_{j=1}^n$ and $(v_j)_{j=1}^n$; i.e. we need to compute

$$\Delta = \min_E \sum_{j=1}^n |Eu_j - v_j|^2$$

To simplify this calculation, first translate O so that its centroid lies at the origin, giving

$$\sum_{j=1}^n u_j = 0$$

Next write E as $Eu = R_\theta u + a$, R_θ denoting a counterclockwise rotation by θ . Then

$$\Delta = \min_{\theta, a} \sum_{j=1}^n |R_\theta u_j + a - v_j|^2 =$$

$$\min_{\theta, a} \left[\sum_{j=1}^n |v_j|^2 + n|a|^2 - 2 \sum_{j=1}^n a \cdot v_j + \sum_{j=1}^n |u_j|^2 + 2 \sum_{j=1}^n a \cdot R_\theta u_j - 2 \sum_{j=1}^n R_\theta u_j \cdot v_j \right]$$

But

$$\sum a \cdot R_\theta u_j = a \cdot R_\theta (\sum u_j) = 0.$$

Hence a and θ appear independently in Δ and we can minimize their contributions separately.

To minimize over \mathbf{a} simply put

$$\mathbf{a} = \frac{1}{n} \sum_{j=1}^n \mathbf{v}_j$$

As to θ , we need to compute

$$\delta = \max_{\theta} \sum_{j=1}^n R_{\theta} \mathbf{u}_j \cdot \mathbf{v}_j$$

Regarding the vectors $\mathbf{u}_j, \mathbf{v}_j$ as complex numbers u_j, v_j , we can rewrite this as

$$\delta = \max_{\theta} \operatorname{Re} \left[\sum_{j=1}^n e^{i\theta} u_j \bar{v}_j \right] = \left| \sum_{j=1}^n u_j \bar{v}_j \right|$$

Altogether this gives

$$\Delta = \sum_{j=1}^n |\mathbf{v}_j|^2 - \frac{1}{n} \left| \sum_{j=1}^n \mathbf{v}_j \right|^2 + \sum_{j=1}^n |u_j|^2 - 2 \left| \sum_{j=1}^n u_j \bar{v}_j \right| \quad (*)$$

$$\Delta = \sum_{j=1}^n |\mathbf{v}_j|^2 - \frac{1}{n} \left| \sum_{j=1}^n \mathbf{v}_j \right|^2 + \sum_{j=1}^n |u_j|^2 - 2 \left(\left| \sum_{j=1}^n \mathbf{u}_j \cdot \mathbf{v}_j \right|^2 + \left| \sum_{j=1}^n \mathbf{u}_j \times \mathbf{v}_j \right|^2 \right)^{\frac{1}{2}},$$

where $\mathbf{u} \times \mathbf{v}$ denotes the (2-dimensional) cross product of the vectors \mathbf{u} and \mathbf{v} . (Note the similarity between (*) and the formula for the best matching between two turning-angle shape descriptors given in the preceding section.)

If O is partially occluded or appears in an unknown orientation, we have to match the sequence $(\mathbf{u}_j)_{j=1}^n$ to each of the contiguous subsequences $(\mathbf{v}_{j+d})_{j=1}^n$ of the (circular) sequence $(\mathbf{v}_j)_{j=1}^m$, for $d = 0, \dots, m-1$. (We assume that $m \geq n$, for otherwise the (partial) periphery of O is too long to match M .)

For each such d (*) becomes

$$\Delta(d) = \sum_{j=d+1}^{d+n} |\mathbf{v}_j|^2 - \frac{1}{n} \left| \sum_{j=d+1}^{d+n} \mathbf{v}_j \right|^2 + \sum_{j=1}^n |u_j|^2 - 2 \left| \sum_{j=1}^n u_j \bar{v}_{j+d} \right|$$

As in the preceding analysis, the minimum of the values $\Delta(d)$, $d = 0, \dots, m-1$, can be found in time $O(m \log m)$, using the fast Fourier transform.

It is interesting to note that the observations made in the last few pages generalize easily to curves in three dimensions, or, more generally, to any situation in which a model curve or surface $\mathbf{u}(\omega)$ depending on one or more parameters ω must be rotated and translated to match a model curve or surface $\mathbf{v}(\omega)$ as well as possible. We need to assume, however, that the matching operation involves no change in parametrization for either of the functions $\mathbf{u}(\omega)$ or $\mathbf{v}(\omega)$.

Suppose more specifically that we are given two descriptor functions $\mathbf{u}(\omega)$, $\mathbf{v}(\omega)$, $\omega \in S$, corresponding respectively to an observed object O and a model object M . We need to find the Euclidean motion E (e.g. of 3-space) which minimizes

$$\Delta = \min_E \int_S |Eu(\omega) - v(\omega)|^2 d\omega$$

As in the 2-D case, we translate O so that its centroid lies at the origin, giving

$$\int_S u(\omega) d\omega = 0$$

Write E as $Eu = Ru + a$, where R is a rotation. Then

$$\Delta = \min_{R, a} \int_S |Ru(\omega) + a - v(\omega)|^2 d\omega =$$

$$\min_{R, a} \left[\int_S |v(\omega)|^2 d\omega + |S||a|^2 - 2 \int_S a \cdot v(\omega) d\omega + \int_S |u(\omega)|^2 d\omega + 2 \int_S a \cdot Ru(\omega) d\omega - 2 \int_S Ru(\omega) \cdot v(\omega) d\omega \right]$$

But

$$\int_S a \cdot Ru(\omega) d\omega = a \cdot R \left(\int_S u(\omega) d\omega \right) = 0.$$

Hence a and R appear independently in Δ and we can minimize their contributions separately.

To minimize over a simply put

$$a = \frac{1}{|S|} \int_S v(\omega) d\omega$$

As to R , we need to compute

$$\delta = \max_R \int_S Ru(\omega) \cdot v(\omega) d\omega$$

To find δ , first calculate the matrix A given by

$$A_{ij} = \int_S u_i(\omega) v_j(\omega) d\omega,$$

(where $i, j = 1, 2, 3$ if we are dealing with a curve or surface in 3-space). In terms of the matrix A we can express δ as

$$\delta = \max_R \text{tr}(RA)$$

To maximize $\text{tr}(RA)$, decompose A as $A = QH$, where $Q = A(A^*A)^{-\frac{1}{2}}$ is a pure rotation and $H = (A^*A)^{\frac{1}{2}}$ is positive definite symmetric. This gives

$$\delta = \max_R \text{tr}(RQH) = \max_R \text{tr}(RH) = \text{tr}(H) = \text{tr}((A^*A)^{\frac{1}{2}})$$

To see this, note that since the trace is invariant under rotation, we can assume

that H is diagonal. But for a diagonal positive definite matrix (λ_i) and a rotation matrix (r_{ij}) the trace of the product $\sum \lambda_i r_{ii}$ can be no larger than $\sum \lambda_i$ and can assume this value only when $r_{ij} = \delta_{ij}$.

Overall, we have

$$\Delta = \int_S |v(\omega)|^2 d\omega - \frac{1}{|S|} \left| \int_S v(\omega) d\omega \right|^2 + \int_S |u(\omega)|^2 d\omega - 2 \operatorname{tr}((A^* A)^{\frac{1}{2}}) \quad (**)$$

showing that the optimal rotated match between O and M can be found in time proportional to that needed to integrate the various functions appearing in (**), i.e. proportional to the number of data points used to discretize the curves or surfaces u and v .

Much as in the 2-D case considered previously, these formulae can be used to match observed 3-D curves parametrized by arc length to similarly parametrized model curves. Matching can be achieved in time $O(n \log n)$ by using the fast Fourier transform, even if the observed curve O is partially obscured. This remark is potentially applicable to matching of 'iso-color' curves on 3-dimensional surfaces.

There are two difficulties in extending the matching technique just described to partially obscured surfaces. The first difficulty is to parametrize O . This point is discussed below; but the obvious parametrization using the centroid that can be used in the unobscured case is not available for partially obscured objects.

A second difficulty involves the computational cost of matching. Suppose that the centroid of O (or some other anchor point common to both O and M) is not known because O is partly obscured. Then we have to match the visible portion of O 's surface against all possible similar portions of M 's surface, and that may force us to iterate over (an appropriate discretization of) the 3-D rotational group R_3 . This means that if we discretize R_3 into n^3 points, and for purpose of integration discretize S into n^2 points, we end up with an $O(n^5)$ matching procedure, far too slow to be useful. What is missing here is an appropriate generalization of the discrete fast Fourier transform algorithm to the case of (a discretized form of) the group O_3 . Even so the complexity can be reduced to $O(n^4 \log n)$ by using the standard fast Fourier transform to handle all n members of O_3 which transform the north pole of S to the same point on S , all simultaneously. To do better than this, a generalization of the fast Fourier transform which can give some rapid way of evaluating integrals on the sphere is needed.

7. Numerical Experiments

The plausibility of the 2-D matching schemes suggested above can be confirmed by simple numerical simulations. For such simulations we begin by generating noisy star-shaped objects, representing hypothetical 'measurements'. These are generated by taking ideal convex objects, and perturbing some specified number of points on each of their sides by adding artificial randomized noise to their distance from the object centroid; this noise ranges between $1 - \text{noise_const}$ and $1 + \text{noise_const}$, where *noise_const* is a specifiable parameter controlling the amount of noise applied. A line of obscuration can also be specified for each simulation run, in which case all points of the polygon lying to the right of this line to be omitted from the simulated measurement. Each generated object *O* is then matched against a collection of ideal convex objects, including the one from which *O* has been generated by applying the above random perturbation.

Each of the two preceding heuristic matching schemes has been simulated. When the matching algorithm finds that two or more ideal objects have nearly the same shape-descriptor distance from a shape-descriptor it reports the error or ambiguity in specific terms.

Easy simulations have been run with the library of convex objects shown in the following figure. The results of these simple experiments are encouraging. Both matching schemes described identify the correct model object in almost all trials. Exceptions occur in cases when an observed object is obscured in a way which made its visible portion similar to a portion of another model object, or when the degree of random noise was high enough to confuse the measured object with a visually similar but different model object (e.g. a circle measured with 20 percent noise may be identified as an oval). The two matching heuristics yield similar results, but in the presence of large quantities of noise the second technique more reliably avoids the grotesque misidentifications that begin to plague the first method. The following figures are generated using the second matching scheme. Note that in each case the matching operation successfully identifies the figure presented to it, from among all the other figures belonging to the small shape library shown in Figure 7, using only those unobscured boundary portions indicated in Figures 8(c) (resp. 8(f)).²

² The authors would like to thank Charles Kim for assistance with the simulations described in this section.

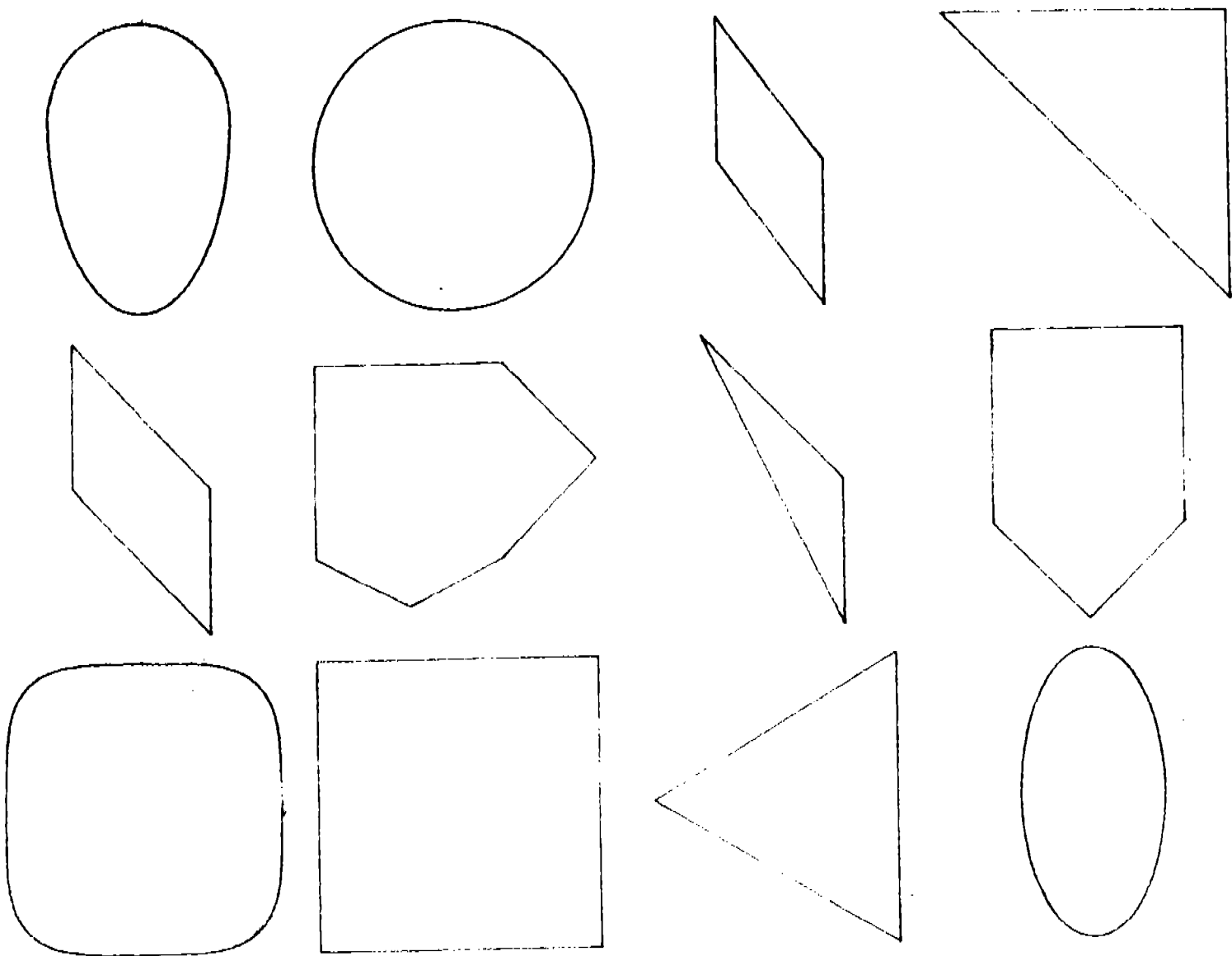


Figure 7: Various figures from library of test figures used in simulations of matching scheme

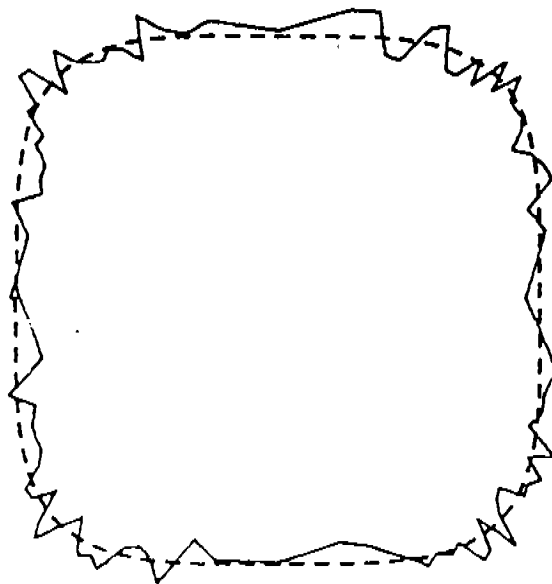


Figure 8(a): A test oval and its roughened form

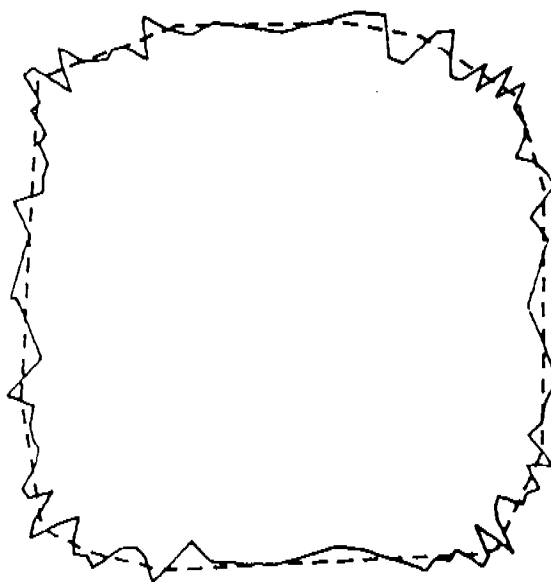


Figure 8(b): Adjusted convex hull of test oval

8. Additional remarks on the turning-angle shape descriptor; some remarks on texture; non-convex regions

Since the derivative of the rotated graph $\eta(s)$ derived from a region's 'turning function' $\theta(s)$ is bounded by 1 in modulus, the Fourier series of $\eta(s)$ converges to $\eta(s)$ with relative rapidity, making it possible to use the first few terms of this series as descriptors of the overall shape of the region. (The adequacy of such an

abbreviated description can be assessed by regenerating the figure from these Fourier coefficients, and then noting what differences with the original region the eye picks out.)

To judge the limitations of this abbreviation, polygon is regular, then the function $\theta(s)$ moves in alternate horizontal and vertical steps which must be equal in size (assuming normalization of the total arc length to 2π). Hence $\eta(s)$ is as shown in the following figure:

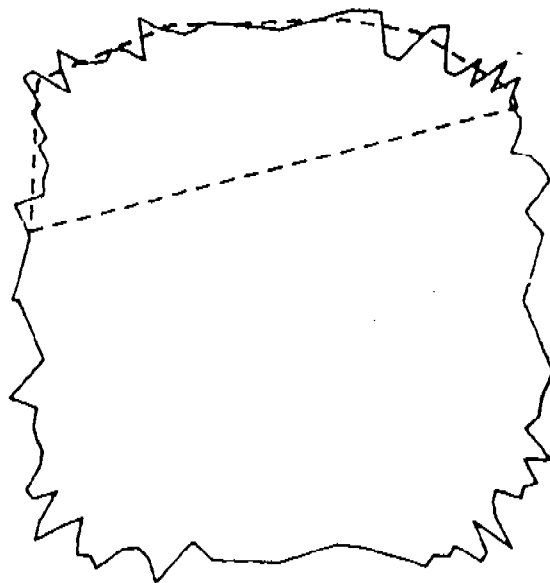


Figure 8(c): Visible portion of test oval hull

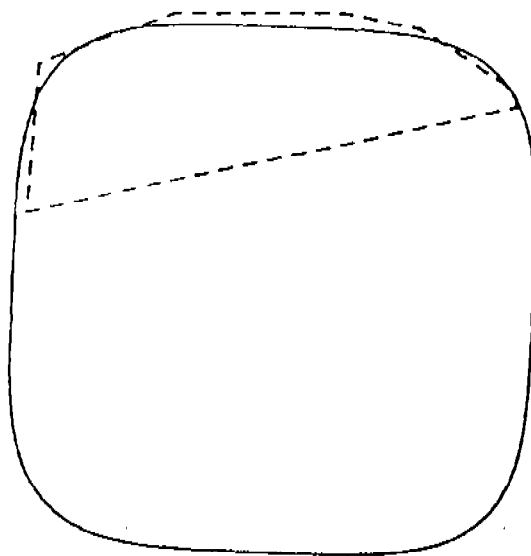


Figure 8(d): Complete match of oval to visible convex hull section

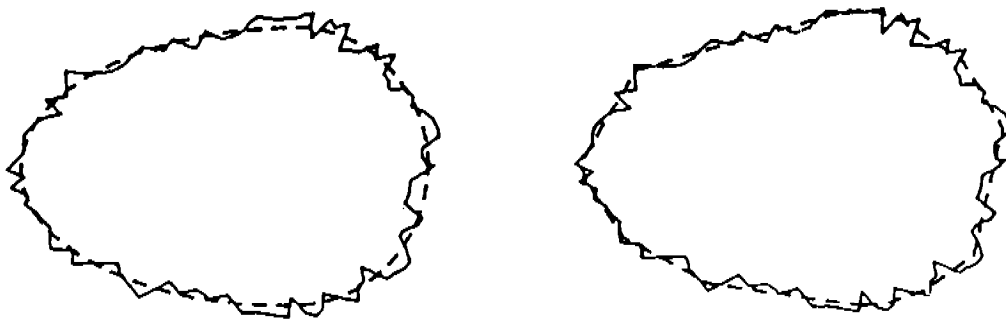


Figure 8(e): Original, roughened form, and adjusted convex hull of a second test oval

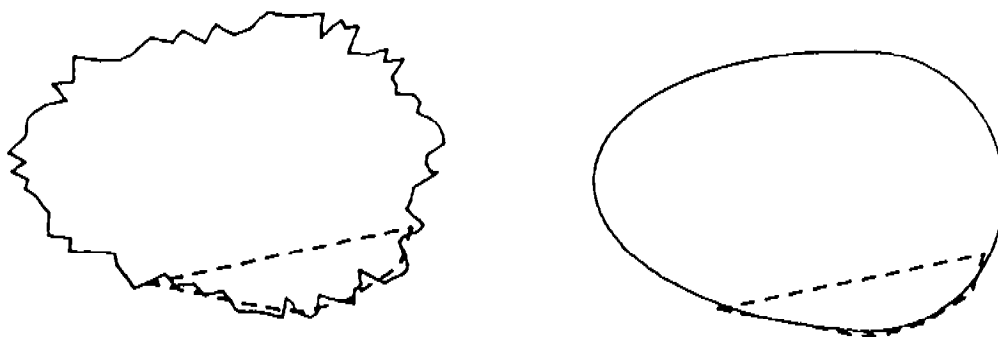


Figure 8(f): Visible portion and computer-generated match for second test oval

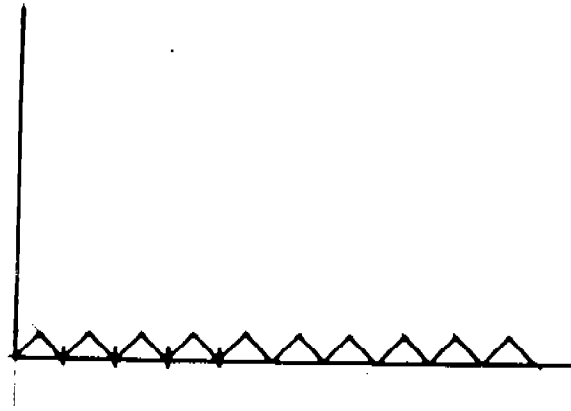


Figure 9: The function $\eta(s)$ for an n -sided regular polygon

A (unit) circle would have a perfectly flat $\eta(s)$ function. This comparison shows that the $\eta(s)$ -function for an n -sided polygon approximates that for a circle very closely in the uniform norm, but that it has a significantly different geometric *texture*. To reflect this fact, a shape-matching scheme would have to apply some operator which will detect texture. The following is one possibility: decompose the function η into parts corresponding to different frequency ranges by applying a disjoint set of band-pass filters. (These can decompose η into its low frequency part, encompassing all frequencies up to a limit F_1 , and into then exponentially expanding frequency ranges F_1 to F_2 , F_2 to F_3 , ... etc.) This gives $\eta(s) = \eta_1(s) + \eta_2(s) + \eta_3(s) + \dots$, where relatively few terms need appear. The low-frequency component can be represented exactly by a few Fourier coefficients, after which each of the few higher-frequency components η_2 , η_3 , ... can be handled as follows: Calculate the variation of each such η_j over the range from 0 to s . This defines a monotone increasing function δ_j ; treat this as previously, i.e. turn it 45° and represent it by its lowest few Fourier terms. The functions δ_j then represent the way that the texture of the original turning function $\theta(s)$ varies from zone to zone along the boundary of the region being analyzed. An approach like this might be able to represent the shape and texture of a convex body adequately using something like 25-35 numerical parameters: e.g. 4 sines and 5 cosines to represent each of η_1 and η_2 , and 2 sines and 3 cosines for η_3 , which will normally be much less significant to the eye.

As in the simpler case noted above, we can assess the adequacy of these descriptors by generating and inspecting the simplest curves which these descriptors fail to distinguish within a variety of examples presented to the analysis system. One obvious shortcoming is that the proposed scheme misses periodicities which the eye can pick out, e.g. it does not distinguish the function η , appearing in the last preceding figure from the function which is different to the eye.

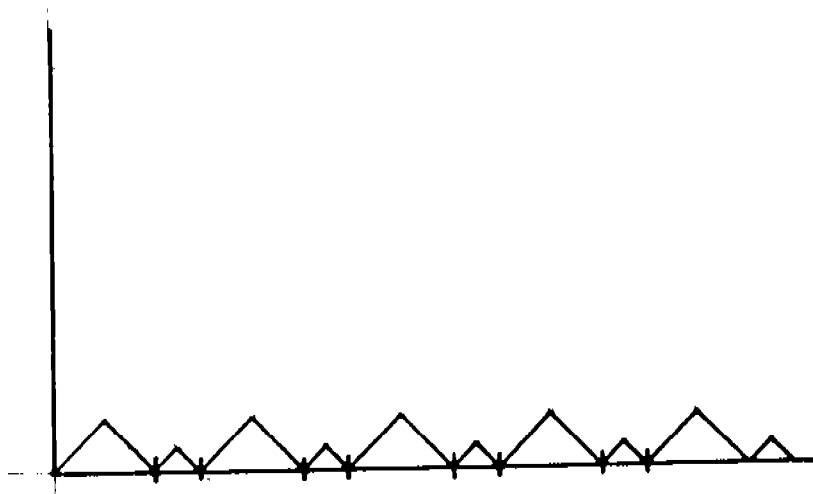


Figure 10: The function $\eta(s)$ for an n -sided but not entirely regular polygon

This remark may be more pessimistic than is justified, since we deal here with small edges on nearly circular polygons with numerous sides; nevertheless, only experiments exhibiting the strengths and weaknesses of the scheme proposed can establish the validity of more refined suggestions.

In concluding this section we note that *non-convex regions* turn out, somewhat surprisingly, to be easier to handle than convex regions. Every concavity is bounded by a single straight side of the convex hull of the body, which can be called the *entrance* to the concavity; if the region is partially obscured, the concavity can be said to have a *correctly visible entrance* if no point of obscuration lies on the opposite side of the entrance from the visible points of the region.

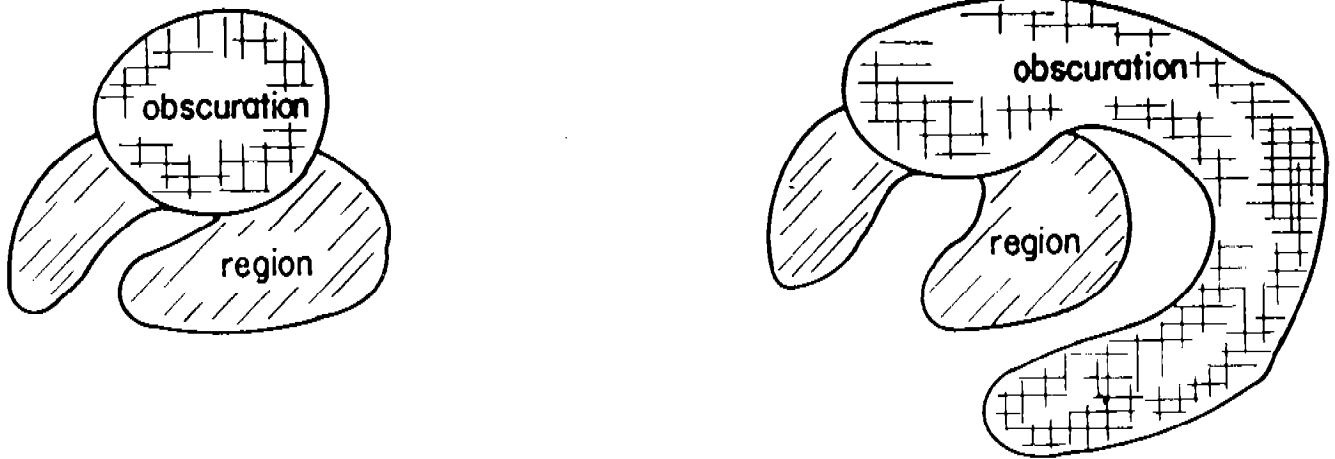


Figure 11: Concavity entrances (The left one is correctly visible, whereas the right one is not.)

A correctly visible concavity entrance identifies a straight side of the convex hull of the region to which it belongs, and since the hull has at least this one straight side we can treat it as 'partially polygonal', i.e. can take the polygon positions in which one such straight side is horizontal as the basis for the search tree used to identify all regions having a concavity with a correctly visible entrance. Moreover, any concavity with a correctly visible entrance can be considered to constitute a region (which may be partially obscured) in its own right; any technique applicable to (partially obscured) regions can be applied recursively to such a concavity. In particular, whenever the whole boundary of a concavity is visible, we can find its centroid and can use this as an anchor point for the region.

9. Three Dimensional Bodies

Having now said a good deal about the 2-D case, we can try to extend our considerations to the more challenging but more significant case of bodies in three dimensions. Let us begin by considering the probing technique, and first its application to the simplest case, that of a convex polyhedron standing on one of its faces. It will generally be easy to find two points p_1, p_2 fixed in the body, e.g. we can form silhouettes of the object as viewed from the x , y , and z directions and use this data to locate the topmost point and the point with largest x -coordinate (note that we need the 3-space locations of both these points). This will be determinable from the three silhouettes unless the topmost (resp. rightmost) edge or

face of the polyhedron is parallel to the xy (resp. yz) plane, in which case appropriate technical adjustments need to be made. Once these points have been found we can logically rotate and translate the body to put these two points in some standard position; then the object is necessarily in one of some finite number of possible positions (the number of these positions being roughly proportional to the number of faces of the polyhedron) and probes using a single-point depth sensor, organized in a 'probe tree' as before, should identify the polyhedron without difficulty.

A similar technique can be used even if the polyhedron does not necessarily stand on a known face. From three silhouettes, we can find the topmost and bottom-most points of the polyhedron, plus the points farthest left and farthest right. These define body position up to one of a finite number of fixed positions. (Similarly, if a horizontal face is topmost or bottommost, this fact, plus the location of the leftmost and rightmost points of the polyhedron, determines its position up to finitely many possibilities.) Then we can probe as before to complete identification and orientation of the polyhedron.

Neither of these techniques depends on polyhedron convexity. Indeed, either technique can be regarded as a way of orienting the convex hull of a non-convex polyhedron. Moreover, as soon as the position of its convex hull is limited to a finite set, the possible positions of the polyhedron itself become equally limited, and probing can be used in the ordinary way to complete its identification.

Next suppose that only part of a polyhedron is visible. If this part includes at least one corner, this corner and an edge running from it can be used to anchor the polyhedron, following which we can apply much the same probing technique as was described for partly obscured polygons.

To extend the probing idea to 3-D objects with smoothly curved boundaries, we need to find, not just one anchor point (as in the 2-D case), but two anchor points (or one anchor point and one 'anchor direction' emerging from it) which have known position relative to the object. Once these points are fixed, the object O is only free to turn about the axis defined by these two points, so that by probing along a circle perpendicular to this axis until contact is made with the body surface, we can restrict O 's possible orientations to a finite set. After this is achieved, the probe tree method can be used to complete the identification of O . Note also that, once a single anchor point p for O has been located, finding a second anchor point q will generally reduce to a relatively easy 2-dimensional problem. If, for example, a sufficiently large portion of O 's surface is visible, we can form the intersection of O with a sphere of appropriate radius about p , and let q be a point whose position in the resulting curve C is fixed; for example, q can be the centroid of C . When too little of O is visible for this approach to work, another possibility is to match C to an appropriate pre-stored model curve using the second of the fast matching techniques described in Section X.

10. 3-D Object Recognition by Shape Descriptors

To apply a shape-descriptor approach we must consider generalizations of the 2-D matching schemes presented above to the 3-D case. When the whole of O is visible, then an advantageous parametrization becomes possible. That is, we can take c to be the centroid of O , and parametrize the points on its boundary by the orientation of the ray connecting them to c . This parametrization is relatively immune to noise. Details are as follows. Let O be a convex 3-D object (if O is an observed object we assume it to be wholly visible). The shape descriptor that we want to use for O is simply a collection of data points $(u(\omega))$ on its boundary, where each point $u(\omega)$ is parametrized by the orientation ω of the ray connecting the centroid c of O with u ; in other words, this shape descriptor is a 3-D vector function defined on the unit sphere S (i.e. a generalized 'coloring' of the sphere).

If O is partially obscured its centroid cannot be determined, but instead we can use any other 'anchor point' having fixed location relative to the visible part of O 's surface as the center for an angle-based parametrization. Once such a fixed parametrization of O 's surface is available, the matching technique described previously may become available; but note the *caveats* concerning efficiency which have been expressed. Note also that the parametrization just outlined is also applicable to the case of nonconvex 3-D surfaces, provided that these surfaces are at least star-shaped with respect to some 'anchor point'.

11. Polyhedron Recognition Using Silhouettes

Given presently available sensors, object silhouettes can be formed more sharply and rapidly than depth images. For this reason, it is worth considering the extent in which the silhouettes of polyhedra can be used to identify them. To this end, the following remarks on silhouettes will be helpful. Suppose that a convex polyhedron P is given a certain orientation in 3-space, and projected upon a plane Q parallel to the xz plane which lies entirely on one side of P . Given any such orientation, one group of P 's faces will be visible from Q , while its other faces will be obscured by the body of P . The boundary between the visible and the invisible portions is a sequence of edges of P , which we will call the *3-silhouette* of P ; the projection onto the xz -plane of the 3-silhouette bounds the ordinary *2-D silhouette* of P , which is always a convex polygon. If we assume that no face of P is orthogonal to the xz -plane, then just one point p of the 3-silhouette projects onto each point q of its 2-D silhouette, and p varies continuously with q . Hence the 3-silhouette is topologically a circle, and therefore divides the surface of P into exactly 2 groups of faces, each of which must be connected. The silhouette of a convex polyhedron P is therefore the projection, on the camera's image plane, of a closed sequence of edges on P .

Suppose we draw the outward-directed normal n to a given face F of P . Then F is visible from Q if n points toward Q , but obscured by the body of P if n points away from Q . To understand how the 3-silhouette of P varies as we rotate P about a vertical axis, it is convenient to project all the normals n to P 's faces F

onto the xy plane. This forms a 'direction diagram' consisting of unit vectors in the xy plane, and then a face is visible from Z if the corresponding projected normal points toward one side, say the negative side, of the y axis, but is invisible otherwise. Thus the edge separating two adjacent faces F_1 and F_2 belongs to P 's 3-silhouette if and only if the projected normal vectors to F_1 and F_2 point into opposite sides of the y axis.

Take an edge of the silhouette and its two extremities u_1, u_2 . These are projections of corresponding polyhedron vertices v_1, v_2 , which therefore lie along two known lines in space. Taking the eye of the camera to be the origin of coordinates, we can therefore write $v_1 = xa_1, v_2 = ya_2$, where a_1, a_2 are known unit vectors. Let u_2u_3 be the next edge of the silhouette. Ignoring exceptional positions, this must correspond to a polyhedron edge v_2v_3 , and again we have $v_3 = za_3$ where a_3 is a known unit vector. It will be noted below that the maximal number of distinct 3-silhouettes that P can have is $O(n^3)$ ($O(n^2)$ if we consider only *isometric* silhouettes), and in fact the maximal number of distinct triplets v_1, v_2, v_3 of adjacent vertices of P in a silhouette is also at most $O(n^3)$ ($O(n^2)$ in the isometric case). Hence (searching as always over a finite number of possibilities) we can suppose that the three distances $D_1 = |v_1v_2|, D_2 = |v_2v_3|, D_3 = |v_3v_1|$ are known. This gives us three quadratic equations for determining the three unknowns x, y, z , namely

$$x^2 + y^2 - 2xy(v_1 \cdot v_2) = D_1^2$$

$$y^2 + z^2 - 2yz(v_2 \cdot v_3) = D_2^2$$

$$z^2 + x^2 - 2xz(v_3 \cdot v_1) = D_3^2$$

These can readily be solved by subtracting suitable multiples of the third equation from the first two, which gives two inhomogeneous quadratic equations for the ratios $\xi = x/z$ and $\eta = y/z$. Thus knowing the positions of these successive vertices on the perimeter of the silhouette determines the polyhedron orientation up to a finite number of possibilities, and hence determines the entire silhouette in the same sense. If the silhouette has four or more vertices we should therefore be able to compare a finite collection of calculated silhouettes with an actual silhouette, and this will often identify the body and its orientation uniquely. Identification becomes even easier if we assume that silhouettes viewed from two slightly different angles are available; we leave it to the reader to work out the details involved.

As usual, the situation is somewhat more favorable if the polyhedron and its silhouette are nonconvex. In such case each vertex of the silhouette which has the property that the silhouette lies on the smaller side of the two silhouette edges forming the silhouette boundary ('convex corners') must correspond to a corner of the polyhedron. Moreover, any segment connecting two such points which does not form part of the silhouette boundary must be the image of an edge ('flying

edge') which connects two corners of the polyhedron but is not an edge of the polyhedron. Often this observation will make it possible to identify silhouette vertices rapidly. Consider, for example, the union of parallelepiped and pyramid shown in the following figure:

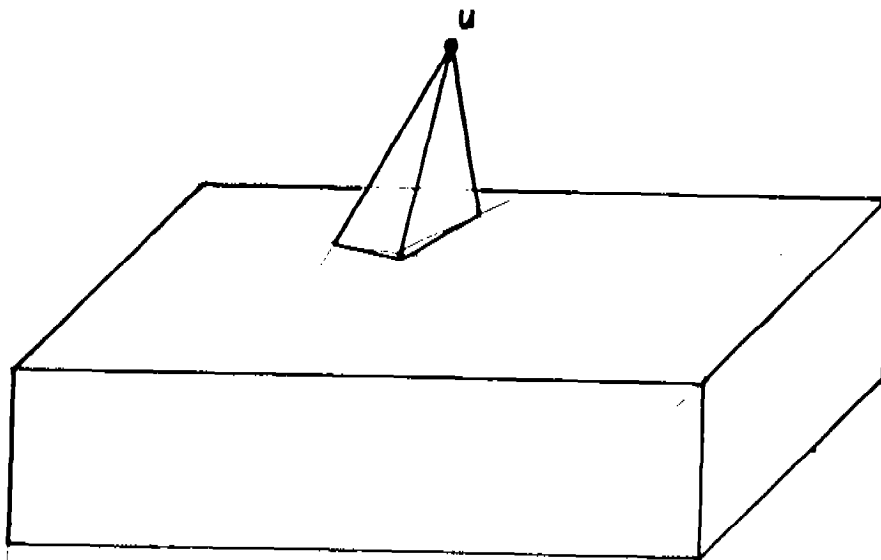


Figure 12: A non-convex polyhedral object

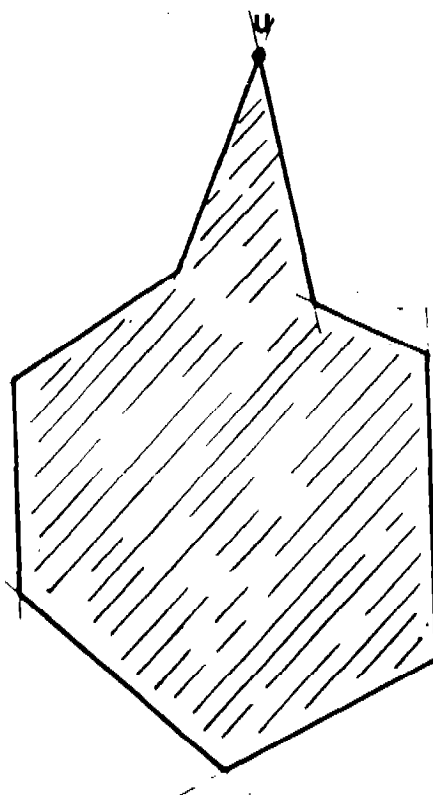


Figure 13: Silhouette of a non-convex polyhedral object

In the silhouette the point u must correspond to v since v is the only polyhedron vertex connected to more than one other vertex by a flying edge.

To show that for any convex polyhedron P there exist at most $O(n^3)$ distinct silhouettes and, in the case of isometric projections, only $O(n^2)$ silhouettes, we can argue as follows. Consider the isometric case first. An isometric silhouette is uniquely determined by the direction in which P is projected, and hence in turn by a point on the unit sphere. As already noted, a silhouette changes its combinatorial structure only when it is projected in a direction v at which some face F of P is seen end-on; in other words, only when v is perpendicular to the normal n_F of F . But for each face F , the locus of orientations v which satisfy this condition is a great circle on the surface of the unit sphere. Since P has n faces, this defines n such great circles which collectively partition the sphere into $O(n^2)$ open regions, inside each of which the combinatorial structure of P 's silhouette remains constant.

Next consider silhouettes seen from an arbitrary viewpoint Z . In this case the combinatorial structure of a silhouette can change only when Z lies on one of the face planes of P . These n planes decompose the 3-D space exterior to P into

$O(n^3)$ regions, within each of which the combinatorial structure of the silhouette remains constant, as asserted.

BIBLIOGRAPHY

- [IS82] K. Ikeuchi and Y. Shirai, A Model-Based Recognition System for Recognition of Machine Parts, Proc. 2nd Annual Nat. Conf. on Artificial Intelligence (1982), pp. 18-21.
- [OS75] M. Oshima and Y. Shirai, Representation of Curved Objects Using Three-Dimensional Information, Proc. 2nd USA-Japan Computer Conf. (1975), pp. 108-112.
- [OS79] M. Oshima and Y. Shirai, A Scene Description Method Using Three-Dimensional Information, Pattern Recognition, v. 11 (1979), pp. 9-17.
- [Sc83] J.T. Schwartz, Structured Light Sensors for 3-D Robot Vision, New York University Technical Report No. 65, Robotics Report No.8, 1983.
- [S79] Y. Shirai, Three-Dimensional Computer Vision, in Computer Vision and Sensor-Based Robots, G. Dodd and L. Rossol (eds.), Plenum Press, N.Y., 1979, pp. 187-205
- [SKOI83] Y. Shirai, K. Koshikawa, M. Oshima, and K. Ikeuchi, A Vision System Based on Three-Dimensional Model, Proc. 1983 Int. Conf. on Advanced Robotics (ICAR83), Tokyo, pp. 139-146.
- [SS71] Y. Shirai and M. Suwa, Recognition of Polyhedrons with a Range Finder, Proc. 2nd Int. Joint Conf. on Artificial Intelligence (1971), pp. 80-87.
- [T83] Technical Arts Corporation, The White Scanner 100 Series, Technical and Sales Brochures, 1983, (Address: 100 Nickerson, Suite j102, Seattle, WA 48109).

CATEGORY LEARNING AND ADAPTIVE PATTERN RECOGNITION: A NEURAL NETWORK MODEL

Gail A. Carpenter[†]

Department of Mathematics, Northeastern University
Boston, Massachusetts 02115

and

Center for Adaptive Systems, Boston University
Boston, Massachusetts 02215

AND

Stephen Grossberg[‡]

Center for Adaptive Systems, Boston University
Boston, Massachusetts 02215

ABSTRACT. A theory is presented of how recognition categories can be learned in response to a temporal stream of input patterns. Interactions between an attentional subsystem and an orienting subsystem enable the network to self-stabilize its learning, without an external teacher, as the code becomes globally self-consistent. Category learning is thus determined by global contextual information in this system. The attentional subsystem learns bottom-up codes and top-down templates, or expectancies. The internal representations formed in this way stabilize themselves against recoding by matching the learned top-down templates against input patterns. This matching process detects structural pattern properties in addition to local feature matches. The top-down templates can also suppress noise in the input patterns, and can subliminally prime the network to anticipate a set of input patterns. Mismatches activate an orienting subsystem, which resets incorrect codes and drives a rapid search for new or more appropriate codes. As the learned code becomes globally self-consistent, the orienting subsystem is automatically disengaged and the memory consolidates. After the recognition categories for a set of input patterns self-stabilize, those patterns directly access their categories without any search or recoding on future recognition trials. A novel pattern exemplar can directly access an established category if it shares invariant properties with the set of familiar exemplars of that category. Several attentional and nonspecific arousal mechanisms modulate the course of search and learning. Three types of attentional mechanism—priming, gain control, and vigilance—are distinguished. Three types of nonspecific arousal are also mechanistically characterized. The nonspecific vigilance process determines how fine the learned categories will be. If vigilance increases due, for example, to a negative reinforcement, then the system automatically searches for and learns finer recognition categories. The learned top-down expectancies become more abstract as the recognition categories become broader. The learned code is a property of network interactions and the entire history of input pattern

[‡] Supported in part by the Air Force Office of Scientific Research (AFOSR 85-0149) and the Army Research Office (ARO DAAG-29-85-K-0095).

[†] Supported in part by the Air Force Office of Scientific Research (AFOSR 85-0149) and the Office of Naval Research (ONR N00014-83-K0337).

Acknowledgements: We wish to thank Cynthia Suchta for her valuable assistance in the preparation of the manuscript.

presentations. The interactions generate emergent rules such as a Weber Law Rule, a 2/3 Rule, and an Associative Decay Rule. No serial programs or algorithmic rule structures are used.

1. Introduction: Self-Organization of Recognition Categories. A fundamental problem of perception and learning concerns the characterization of how recognition categories emerge as a function of experience. When such categories spontaneously emerge through an individual's interaction with an environment, the processes are said to undergo *self-organization* [1]. A theory of how recognition categories can self-organize is outlined in this report, which summarizes the model's design and mathematical analysis, developed in other articles [2-4]. In those articles, the *adaptive resonance theory* is also related to recent data about evoked potentials and about amnesias due to malfunction of medial temporal brain structures. Results of evoked potential and clinical studies suggest which macroscopic brain structures could carry out the theoretical dynamics. The theory also specifies microscopic neural dynamics, with local processes obeying membrane equations (Appendix).

We focus herein upon principles and mechanisms that are capable of self-organizing stable recognition codes in response to arbitrary temporal sequences of input patterns. These principles and mechanisms lead to the design of a neural network whose parameters can be specialized for applications to particular problem domains, such as speech and vision. In these domains, preprocessing stages prepare environmental inputs for the self-organizing category formation and recognition system. Work on speech and language preprocessing has characterized those stages after which such a self-organizing recognition system can build up codes for phonemes, syllables, and words [5-7]. Work on form and color preprocessing has characterized those stages after which such a self-organizing recognition system can build up codes for visual object recognition [8,9].

2. Bottom-Up Adaptive Filtering and Contrast-Enhancement in Short Term Memory. We now introduce in a qualitative way the main mechanisms of the theory. We do so by considering the typical network reactions to a single input pattern I within a temporal stream of input patterns. Each input pattern may be the output pattern of a preprocessing stage. The input pattern I is received at the stage F_1 of an *attentional subsystem*. Pattern I is transformed into a pattern X of activation across the nodes of F_1 (Figure 1). The transformed pattern X represents a pattern in short term memory (STM). In F_1 each node whose activity is sufficiently large generates excitatory signals along pathways to target nodes at the next processing stage F_2 . A pattern X of STM activities across F_1 hereby elicits a pattern S of output signals from F_1 . When a signal from a node in F_1 is carried along a pathway to F_2 , the signal is multiplied, or *gated*, by the pathway's long term memory (LTM) trace. The LTM gated signal (i.e., signal times LTM trace), not the signal alone, reaches the target node. Each target node sums up all of its LTM gated signals. In this way, pattern S generates a pattern T of LTM-gated and summed input signals to F_2 (Figure 2a). The transformation from S to T is called an *adaptive filter*.

The input pattern T to F_2 is quickly transformed by interactions among the nodes of F_2 . These interactions contrast-enhance the input pattern T. The resulting pattern of activation across F_2 is a new pattern Y. The contrast-enhanced pattern Y, rather than the input pattern T, is stored in STM by F_2 .

A special case of this contrast-enhancement process, in which F_2 chooses the node which receives the largest input, is here considered. The chosen node is the only one that can store activity in STM. In more general versions of the theory, the contrast enhancing transformation from T to Y enables more than one node at a time to be active in STM. Such transformations are designed to simultaneously represent in STM many subsets, or

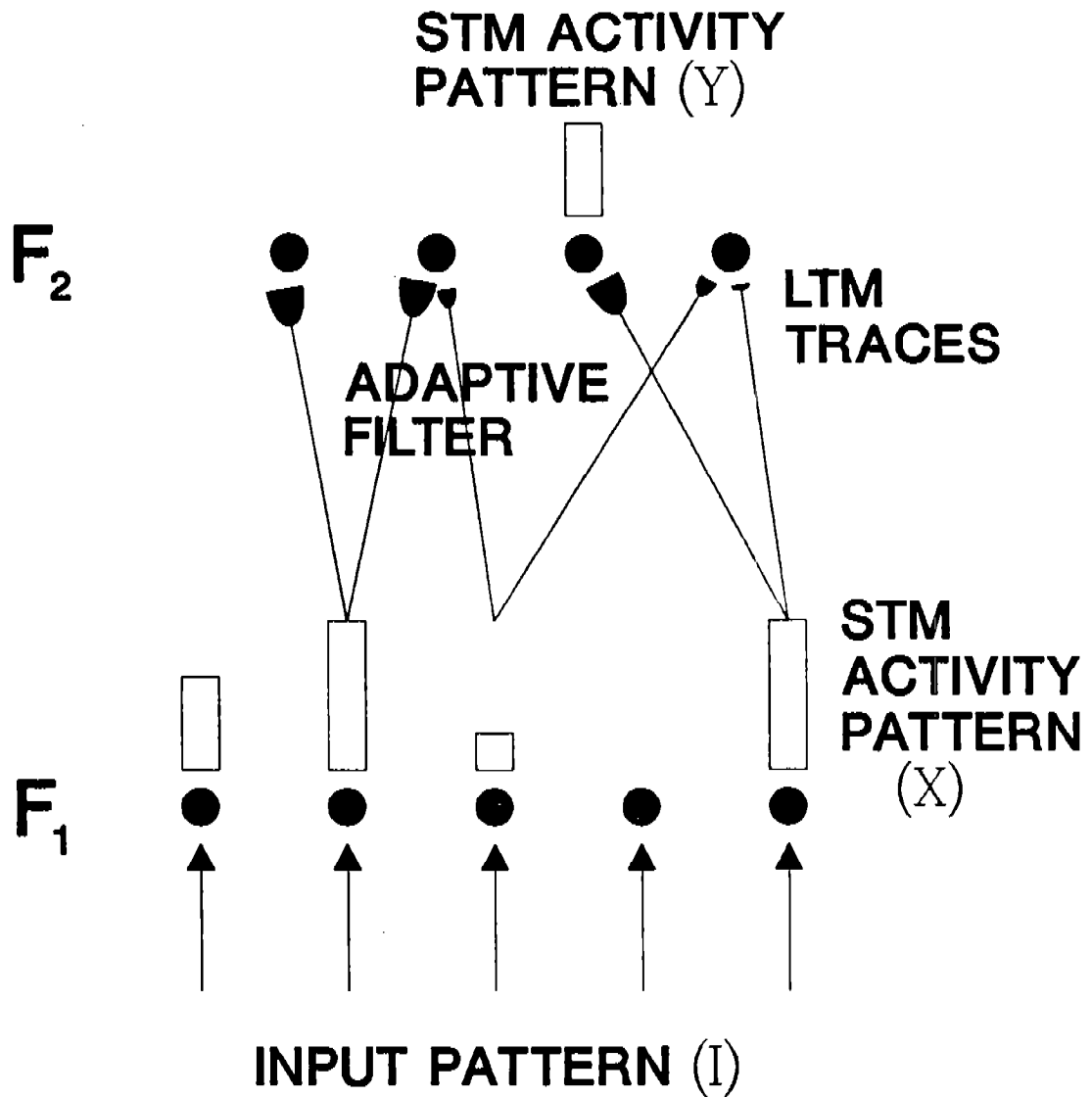


Figure 1. Stages of bottom-up activation: The input pattern I generates a pattern of STM activation X across F_1 . Sufficiently active F_1 nodes emit bottom-up signals to F_2 . This signal pattern S is gated by long term memory (LTM) traces within the $F_1 \rightarrow F_2$ pathways. The LTM-gated signals are summed before activating their target nodes in F_2 . This LTM-gated and summed signal pattern T generates a pattern of activation Y across F_2 .

groupings. of an input pattern [6,10]. When F_2 is designed to make a choice in STM, it selects that global grouping of the input pattern which is preferred by the adaptive filter. This process automatically enables the network to partition all the input patterns which are received by F_1 into disjoint sets of recognition categories, each corresponding to a particular node in F_2 .

Only those nodes of F_2 which maintain stored activity in STM can elicit new learning at contiguous LTM traces. Whereas all the LTM traces in the adaptive filter, and thus all learned past experiences of the network, are used to determine recognition via the transformation $I \rightarrow X \rightarrow S \rightarrow T \rightarrow Y$, only those LTM traces whose STM activities in F_2 survive the contrast-enhancement process can learn in response to the activity pattern X .

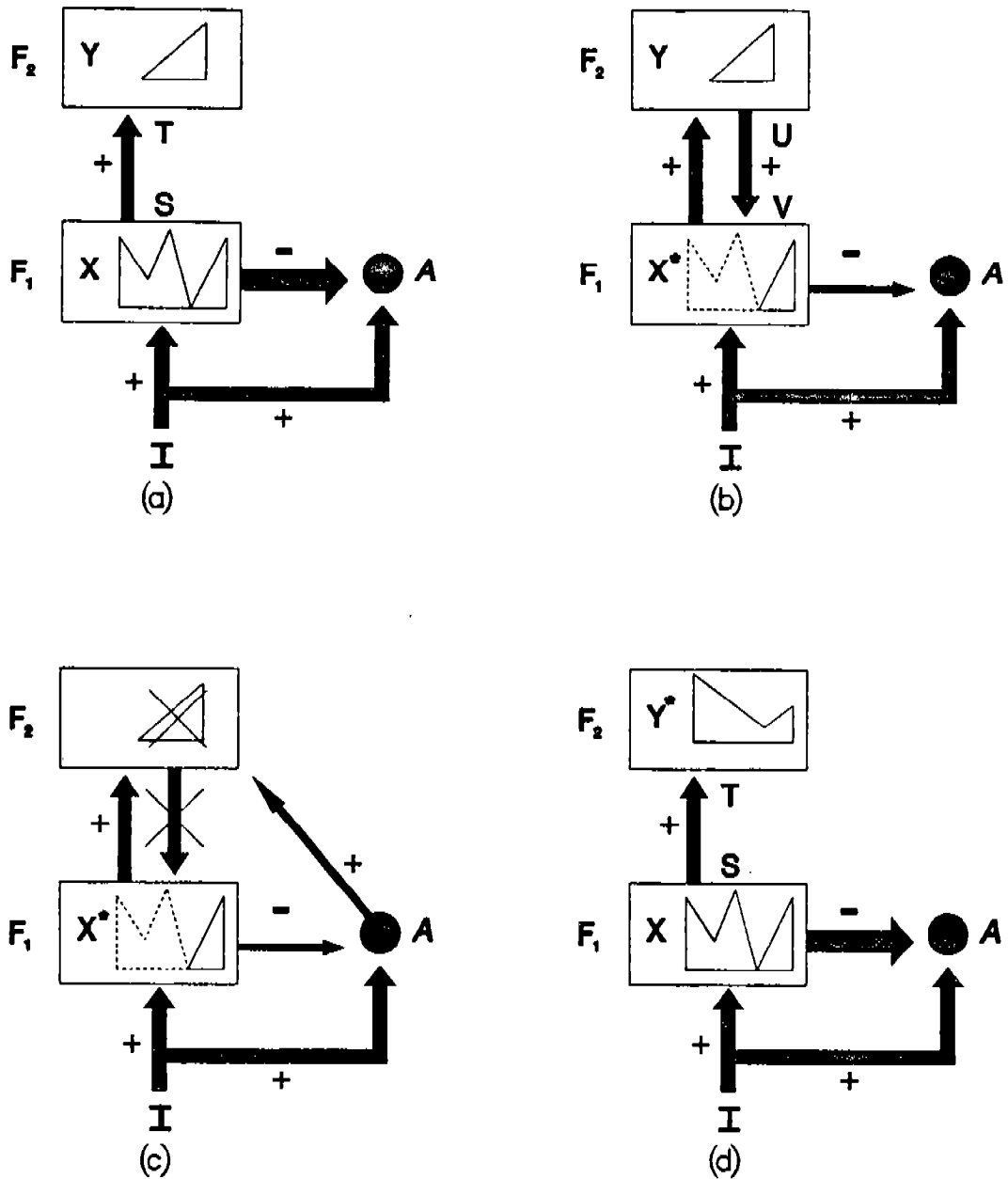


Figure 2. Search for a correct F₂ code: (a) The input pattern I generates the specific STM activity pattern X at F₁ as it nonspecifically activates A. Pattern X both inhibits A and generates the output signal pattern S. Signal pattern S is transformed into the input pattern T, which activates the STM pattern Y across F₂. (b) Pattern Y generates the top-down signal pattern U which is transformed into the template pattern V. If V mismatches I at F₁, then a new STM activity pattern X* is generated at F₁. The reduction in total STM activity which occurs when X is transformed into X* causes a decrease in the total inhibition from F₁ to A. (c) Then the input-driven activation of A can release a nonspecific arousal wave to F₂, which resets the STM pattern Y at F₂. (d) After Y is inhibited, its top-down template is eliminated, and X can be reinstated at F₁. Now X once again generates input pattern T to F₂, but since Y remains inhibited T can activate a different STM pattern Y* at F₂. If the top-down template due to Y* also mismatches I at F₁, then the rapid search for an appropriate F₂ code continues.

The bottom-up STM transformation $I \rightarrow X \rightarrow S \rightarrow T \rightarrow Y$ is not the only process that regulates network learning. In the absence of top-down processing, the LTM traces within the adaptive filter $S \rightarrow T$ (Figure 2a) can respond to certain sequences of input patterns by being ceaselessly recoded in such a way that individual events are never eventually encoded by a single category no matter how many times they are presented. An infinite class of examples in which temporally unstable codes evolve is described in Section 7. It was the instability of bottom-up adaptive coding that led Grossberg [11,12] to introduce the adaptive resonance theory.

In the adaptive resonance theory, a matching process at F_1 exists whereby learned top-down expectancies, or templates, from F_2 to F_1 are compared with the bottom-up input pattern to F_1 . This matching process stabilizes the learning that emerges in response to an arbitrary input environment. The constraints that follow from the need to stabilize learning enable us to choose among the many possible versions of top-down template matching and STM processes. These learning constraints upon the adaptive resonance top-down design have enabled the theory to explain data from visual and auditory information processing experiments in which learning has not been a manipulated variable [4,6,7]. These mechanisms have now been developed into a rigorously characterized learning system whose properties have been quantitatively analysed [2,3]. This analysis has revealed new design constraints within the adaptive resonance theory. The system that we will describe for learned categorical recognition is one outcome of this analysis.

Figure 3 summarizes the total network architecture. It includes modulatory processes, such as attentional gain control, which regulate matching within F_1 , as well as modulatory processes, such as orienting arousal, which regulate reset within F_2 . Figure 3 also includes an attentional gain control process at F_2 . Such a process enables offset of the input pattern to terminate all STM activity within the attentional subsystem in preparation for the next input pattern. In this example, STM storage can persist after the input pattern terminates only if an internally generated or intermodality input source maintains the activity of the attentional gain control system.

3. Top-Down Template Matching and Stabilization of Code Learning. We now begin to consider how top-down template matching can stabilize code learning. In order to do so, top-down template matching at F_1 must be able to prevent learning at bottom-up LTM traces whose contiguous F_2 nodes are only momentarily activated in STM. This ability depends upon the different rates at which STM activities and LTM traces can change. The STM transformation $I \rightarrow X \rightarrow S \rightarrow T \rightarrow Y$ takes place very quickly. By "very quickly" we mean much more quickly than the rate at which the LTM traces in the adaptive filter $S \rightarrow T$ can change. As soon as the bottom-up STM transformation $X \rightarrow Y$ takes place, the STM activities Y in F_2 elicit a top-down excitatory signal pattern U back to F_1 . Only sufficiently large STM activities in Y elicit signals in U along the feedback pathways $F_2 \rightarrow F_1$.

As in the bottom-up adaptive filter, the top-down signals U are also gated by LTM traces before the LTM-gated signals are summed at F_1 nodes. The pattern U of output signals from F_2 hereby generates a pattern V of LTM-gated and summed input signals to F_1 . The transformation from U to V is thus also an adaptive filter. The pattern V is called a *top-down template*, or *learned expectation* (Figure 2b).

Two sources of input now perturb F_1 : the bottom-up input pattern I which gave rise to the original activity pattern X , and the top-down template pattern V that resulted from activating X . The activity pattern X^* across F_1 that is induced by I and V taken together is typically different from the activity pattern X that was previously induced by I alone. In particular, F_1 acts to match V against I . The result of this matching process determines the future course of learning and recognition by the network.

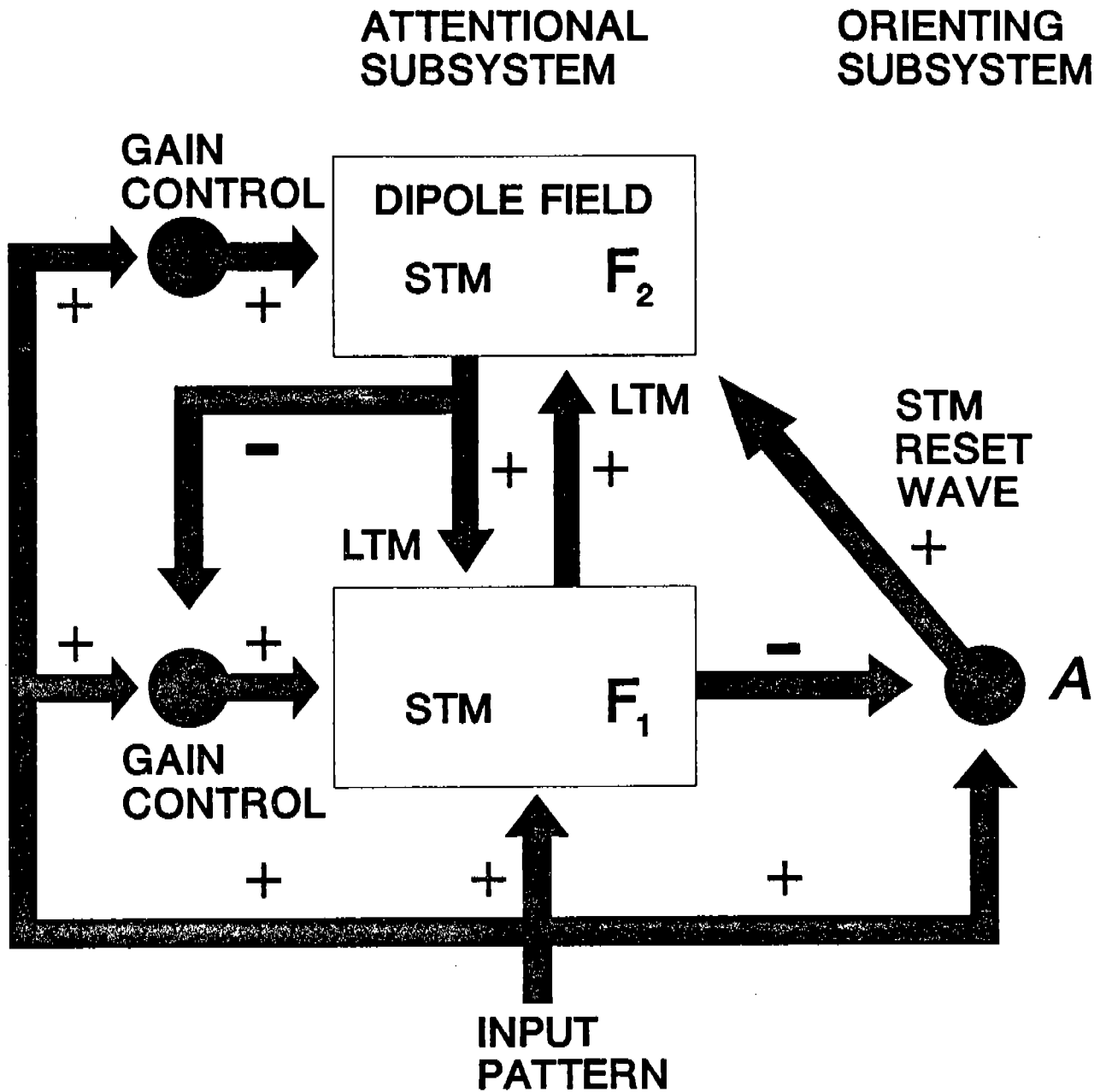


Figure 3. Anatomy of the attentional-orienting system: This figure describes all the interactions of the model without regard to which components are active at any given time.

The entire activation sequence

$$I \rightarrow X \rightarrow S \rightarrow T \rightarrow Y \rightarrow U \rightarrow V \rightarrow X^* \quad (1)$$

takes place very quickly relative to the rate with which the LTM traces in either the bottom-up adaptive filter $S \rightarrow T$ or the top-down adaptive filter $U \rightarrow V$ can change. Even though none of the LTM traces changes during such a short time, their prior learning strongly influences the STM patterns Y and X^* that evolve within the network. We now

discuss how a match or mismatch of I and V at F_1 regulates the course of learning in response to the pattern I.

4. Interactions between Attentional and Orienting Subsystems: STM Reset and Search. This section outlines how a mismatch at F_1 regulates the learning process. With this general scheme in mind, we will be able to consider details of how bottom-up filters and top-down templates are learned and how matching takes place.

Level F_1 can compute a match or mismatch between a bottom-up input pattern I and a top-down template pattern V, but it cannot compute which STM pattern Y across F_2 generated the template pattern V. Thus the outcome of matching at F_1 must have a nonspecific effect upon F_2 that can potentially influence all of the F_2 nodes, any one of which may have read-out V. The internal organization of F_2 must be the agent whereby this nonspecific event, which we call a *reset wave*, selectively alters the stored STM activity pattern Y. The reset wave is one of the three types of nonspecific arousal that exist within the network. In particular, we suggest that a mismatch of I and V within F_1 generates a nonspecific arousal burst that inhibits the active population in F_2 which read-out V. In this way, an erroneous STM representation at F_2 is quickly eliminated before any LTM traces can encode this error.

The attentional subsystem works together with an *orienting subsystem* to carry out these interactions. All learning takes place within the attentional subsystem. All matches and mismatches are computed within the attentional subsystem. The orienting subsystem is the source of the nonspecific arousal bursts that reset STM within level F_2 of the attentional subsystem. The outcome of matching within F_1 determines whether or not such an arousal burst will be generated by the orienting subsystem. Thus the orienting system mediates reset of F_2 due to mismatches within F_1 .

Figure 2 depicts a typical interaction between the attentional subsystem and the orienting subsystem. In Figure 2a, an input pattern I instates an STM activity pattern X across F_1 . The input pattern I also excites the orienting population A, but pattern X at F_1 inhibits A before it can generate an output signal.

Activity pattern X also generates an output pattern S which, via the bottom-up adaptive filter, instates an STM activity pattern Y across F_2 . In Figure 2b, pattern Y reads a top-down template pattern V into F_1 . Template V mismatches input I, thereby significantly inhibiting STM activity across F_1 . The amount by which activity in X is attenuated to generate X^* depends upon how much of the input pattern I is encoded within the template pattern V.

When a mismatch attenuates STM activity across F_1 , this activity no longer prevents the arousal source A from firing. Figure 2c depicts how disinhibition of A releases a nonspecific arousal burst to F_2 . This arousal burst, in turn, selectively inhibits the active population in F_2 . This inhibition is long-lasting. One physiological design for F_2 processing which has these necessary properties is a *dipole field* [4.13]. A dipole field consists of opponent processing channels which are gated by habituating chemical transmitters. A nonspecific arousal burst induces selective and enduring inhibition within a dipole field. In Figure 2c, inhibition of Y leads to inhibition of the top-down template V, and thereby terminates the mismatch between I and V. Input pattern I can thus reinstate the activity pattern X across F_1 , which again generates the output pattern S from F_1 and the input pattern T to F_2 . Due to the enduring inhibition at F_2 , the input pattern T can no longer activate the same pattern Y at F_2 . A new pattern Y^* is thus generated at F_2 by I (Figure 2d). Despite the fact that some F_2 nodes may remain inhibited by the STM reset property, the new pattern Y^* may encode large STM activities. This is because level F_2 is designed so that its total suprathreshold activity remains approximately constant, or normalized, despite the fact that some of its nodes may remain inhibited by the STM reset mechanism. This property is related to the limited capacity of STM. A physiological process capable

of achieving the STM normalization property can be based upon on-center off-surround interactions among cells obeying membrane equations [4.14].

The new activity pattern Y^* reads-out a new top-down template pattern V^* . If a mismatch again occurs at F_1 , the orienting subsystem is again engaged, thereby leading to another arousal-mediated reset of STM at F_2 . In this way, a rapid series of STM matching and reset events may occur. Such an STM matching and reset series controls the system's search of LTM by sequentially engaging the novelty-sensitive orienting subsystem. Although STM is reset sequentially in time, the mechanisms which control the LTM search are all parallel network interactions, rather than serial algorithms. Such a parallel search scheme is necessary in a system whose LTM codes do not exist *a priori*. In general, the spatial configuration of codes in such a system depends upon both the system's initial configuration and its unique learning history. Consequently, no prewired serial algorithm could possibly anticipate an efficient order of search.

The mismatch-mediated search of LTM ends when an STM pattern across F_2 reads-out a top-down template which either matches I to the degree of accuracy required by the level of attentional vigilance, or has not yet undergone any prior learning. In the latter case, a new recognition category is established as a bottom-up code and top-down template are learned.

We now begin to consider details of the bottom-up/top-down matching process across F_1 . The nature of this matching process is clarified by a consideration of how F_1 distinguishes between activation by bottom-up inputs and top-down templates.

5. Attentional Gain Control and Attentional Priming. The importance of the distinction between bottom-up and top-down processing becomes evident when one observes that the same top-down template matching process which stabilizes learning is also a mechanism of attentional priming. Consider, for example, a situation in which F_2 is activated by a level other than F_1 before F_1 is itself activated. In such a situation, F_2 can generate a top-down template V to F_1 . The level F_1 is then primed, or ready, to receive a bottom-up input that may or may not match the active expectancy. Level F_1 can be primed to receive a bottom-up input without necessarily eliciting suprathreshold output signals in response to the priming expectancy. If this were not possible, then every priming event would lead to suprathreshold consequences. Such a property would prevent subliminal anticipation of a future event.

On the other hand, an input pattern I must be able to generate a suprathreshold activity pattern X even if no top-down expectancy is active across F_1 (Figure 2). How does F_1 know that it should generate a suprathreshold reaction to a bottom-up input pattern but not to a top-down input pattern? In both cases, an input pattern stimulates F_1 cells. Some auxiliary mechanism must exist to distinguish between bottom-up and top-down inputs. We call this auxiliary mechanism *attentional gain control* to distinguish it from *attentional priming* by the top-down template itself. The attentional priming mechanism delivers *specific* template patterns to F_1 . The attentional gain control mechanism has a *nonspecific* effect on the sensitivity with which F_1 responds to the template pattern, as well as to other patterns received by F_1 . Attentional gain control is one of the three types of nonspecific arousal that exist within the network. With the addition of attentional gain control, we can explain qualitatively how F_1 can tell the difference between bottom-up and top-down signal patterns.

The need to dissociate attentional priming from attentional gain control can also be seen from the fact that top-down priming events do not lead necessarily to subliminal reactions at F_1 . Under certain circumstances, top-down expectancies can lead to suprathreshold consequences. We can, for example, experience internal conversations or images at will. Thus there exists a difference between the read-out of a top-down template, which is a mechanism of attentional priming, and the translation of this operation into suprathreshold

signals due to attentional gain control. An "act of will" can amplify attentional gain control signals to elicit a suprathreshold reaction at F_1 in response to an attentional priming pattern from F_2 .

Figure 4 depicts one possible scheme whereby supraliminal reactions to bottom-up signals, subliminal reactions to top-down signals, and supraliminal reactions to matched bottom-up and top-down signals can be achieved. Figure 4d shows, in addition, how competitive interactions across modalities can prevent F_1 from generating a supraliminal reaction to bottom-up signals, as when attention shifts from one modality to another.

6. Matching: The 2/3 Rule. We can now outline the matching and coding properties that are used to generate learning of self-stabilizing recognition categories. Two different types of properties need to be articulated: the bottom-up coding properties which determine the order of search, and the top-down matching properties which determine whether an STM reset event will be elicited. Order of search is determined entirely by properties of the attentional subsystem. The choice between STM reset and STM resonance is dependent upon whether or not the orienting subsystem will generate a reset wave. This computation is based on inputs received by the orienting subsystem from both the bottom-up input pattern I and the STM pattern which F_1 computes within the attentional subsystem (Figure 2). Both the order of search and the choice between reset and resonance are sensitive to the matched patterns *as a whole*. This global sensitivity is key to the design of a single system capable of matching patterns in which the number of coded features, or details, may vary greatly. Such global context-sensitivity is needed to determine whether a fixed amount of mismatch should be treated as functional noise, or as an event capable of eliciting search for a different category. For example, one or two details may be sufficient to differentiate two small but functionally distinct patterns, whereas the same details, embedded in a large, complex pattern may be quite irrelevant.

We first discuss the properties which determine the order of search. Network interactions which control search order can be described in terms of three rules: the 2/3 Rule, the Weber Law Rule, and the Associative Decay Rule.

The 2/3 Rule follows naturally from the distinction between attentional gain control and attentional priming. It says that two out of three signal sources must activate an F_1 node in order for that node to generate suprathreshold output signals. In Figure 4a, for example, during bottom-up processing, a suprathreshold node in F_1 is one which receives a specific input from the input pattern I and a nonspecific attentional gain control signal. All other nodes in F_1 receive only the nonspecific gain control signal. Since these cells receive inputs from only one pathway they do not fire.

In Figure 4b, during top-down processing, or priming, some nodes in F_1 receive a template signal from F_2 , whereas other nodes receive no signal whatsoever. All the nodes of F_1 receive inputs from at most one of their three possible input sources. Hence no cells in F_1 are supraliminally activated by a top-down template.

During simultaneous bottom-up and top-down signalling, the attentional gain control signal is inhibited by the top-down channel (Figure 4c). Despite this fact, some nodes of F_1 may receive sufficiently large inputs from both the bottom-up and the top-down signal patterns to generate suprathreshold outputs. Other nodes may receive inputs from the top-down template pattern or the bottom-up input pattern, but not both. These nodes receive signals from only one of their possible sources, hence do not fire. Cells which receive no inputs do not fire either. Thus only cells that are conjointly activated by the bottom-up input and the top-down template can fire when a top-down template is active. The 2/3 Rule clarifies the apparently paradoxical process whereby the addition of top-down excitatory inputs to F_1 can lead to an overall decrease in F_1 's STM activity (Figures 2a and 2b).

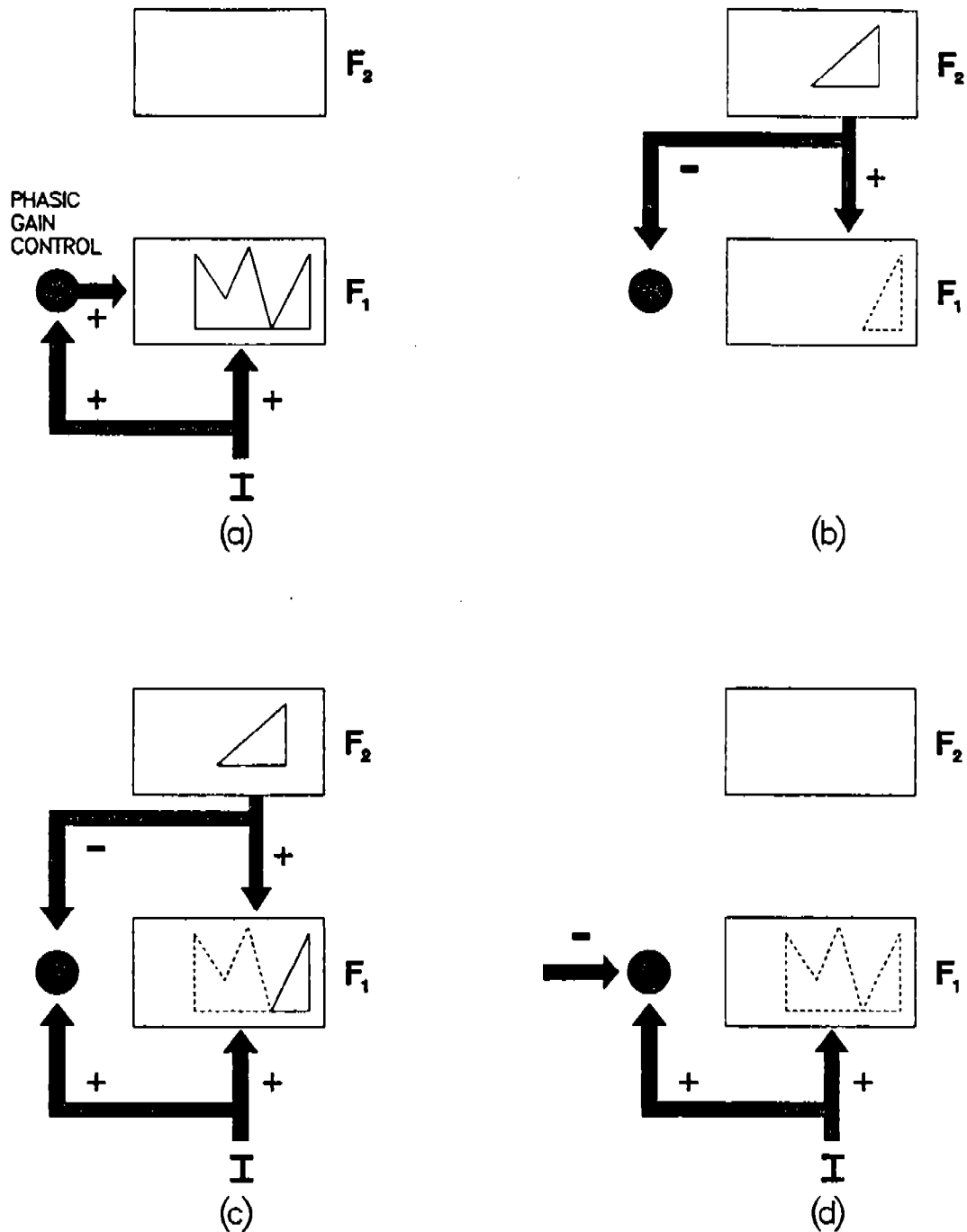


Figure 4. Matching by 2/3 Rule: (a) In this example, nonspecific attentional gain control signals are phasically activated by the bottom-up input. In this network, the bottom-up input arouses two different nonspecific channels: the attentional gain control channel and the orienting subsystem. Only F_1 cells that receive bottom-up inputs and gain control signals can become supraliminally active. (b) A top-down template from F_2 inhibits the attentional gain control source as it subliminally primes target F_1 cells. (c) When a bottom-up input pattern and a top-down template are simultaneously active, only those F_1 cells that receive inputs from both sources can become supraliminally active, since the gain control source is inhibited. (d) Intermodality inhibition can shut off the gain control source and thereby prevent a bottom-up input from supraliminally activating F_1 .

7. Example of Code Instability. We now illustrate the importance of the 2/3 Rule by describing how its absence can lead to a temporally unstable code. In the simplest type of code instability example, the code becomes unstable because neither top-down template nor reset mechanisms exist [11]. Then, in response to certain input sequences that are repeated through time, a given input pattern can be ceaselessly recoded into more than one category. In the example that we will now describe, the top-down template signals are active and the reset mechanism is functional. However, the inhibitory top-down attentional gain control signals (Figures 3 and 4c) are chosen too small for the 2/3 Rule to hold at F_1 . We show also that a larger choice of attentional gain control signals restores code stability by reinstating the 2/3 Rule. These simulations also illustrate three other points: how a novel exemplar can directly access a previously established category; how the category in which a given exemplar is coded can be influenced by the categories which form to encode very different exemplars; and how the network responds to exemplars as coherent groupings of features, rather than to isolated feature matches or mismatches.

Figure 5a summarizes a computer simulation of unstable code learning. Figure 5b summarizes a computer simulation that illustrates how reinstatement of the 2/3 Rule can stabilize code learning.

The first column of Figure 5a describes the four input patterns that were used in the simulation. These input patterns are labeled A, B, C, and D. Patterns B, C, and D are all subsets of A. The relationships among the inputs that make the simulation work are as follows:

Code Instability Example

$$D \subset C \subset A. \quad (2)$$

$$B \subset A. \quad (3)$$

$$B \cap C = \emptyset. \quad (4)$$

$$|D| < |B| < |C|. \quad (5)$$

These results thus provide infinitely many examples in which an alphabet of just four input patterns cannot be stably coded without the 2/3 Rule. The numbers 1, 2, 3, ... listed in the second column itemize the presentation order. The third column, labeled BU for Bottom-Up, describes the input pattern that was presented on each trial. In both Figures 5a and 5b, the input patterns were periodically presented in the order ABCAD.

Each of the Top-Down Template columns in Figure 5 corresponds to a different node in F_2 , with column 1 corresponding to node v_1 , column 2 corresponding to node v_2 , and so on. Each row summarizes the network response to its input pattern. The symbol RES, which stands for *resonance*, designates the node in F_2 which codes the input pattern on that trial. For example, v_2 codes pattern C on trial 3, and v_1 codes pattern B on trial 7. The patterns in a given row describe the templates after learning has occurred on that trial.

In Figure 5a, input pattern A is periodically recoded: On trial 1, it is coded by v_1 ; on trial 4, it is coded by v_2 ; on trial 6, it is coded by v_1 ; on trial 9, it is coded by v_2 . This alternation in the nodes v_1 and v_2 which code pattern A repeats indefinitely.

Violation of the 2/3 Rule occurs on trials 4, 6, 8, 9, and so on. This violation is illustrated by comparing the template of v_2 on trials 3 and 4. On trial 3, the template of v_2 is coded by pattern C, which is a subset of pattern A. On trial 4, pattern A is presented and directly activates node v_2 . Because the 2/3 Rule does not hold, pattern A remains supraliminal in F_1 even after the subset template C is read-out from v_2 . Thus no search is elicited by the mismatch of pattern A and its subset template C. Consequently the template of v_2 is recoded from pattern C to its superset pattern A.

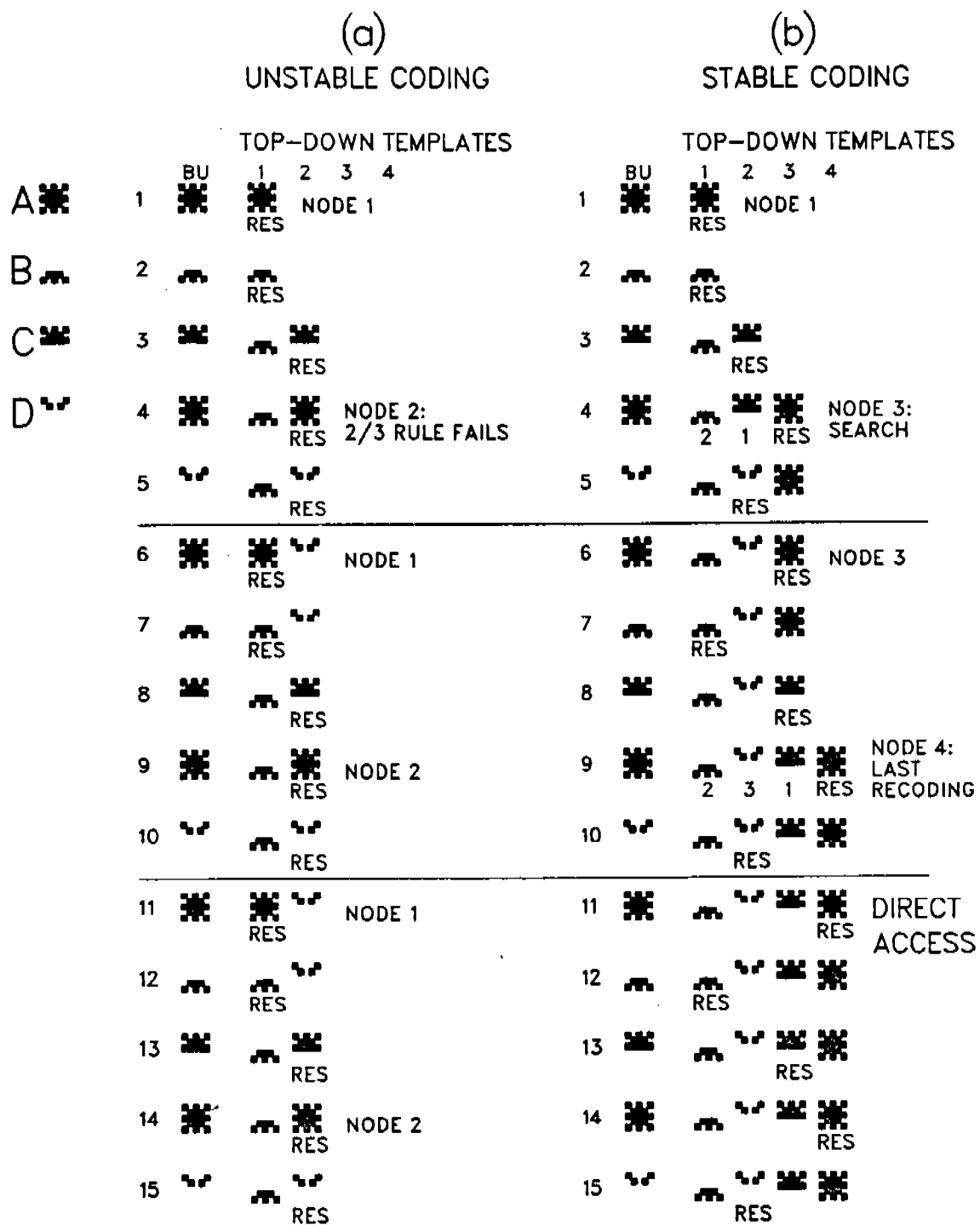


Figure 5. Stabilization of categorical learning by the 2/3 Rule: In both (a) and (b), four input patterns A, B, C, and D are presented repeatedly in the list order ABCAD. In (a), the 2/3 Rule is violated because the top-down inhibitory gain control mechanism is weak (Figure 4c). Pattern A is periodically coded by v_1 and v_2 . It is never coded by a single stable category. In (b), the 2/3 Rule is restored by strengthening the top-down inhibitory gain control mechanism. After some initial recoding during the first two presentations of ABCAD, all patterns directly access distinct stable categories.

In Figure 5b, by contrast, the 2/3 Rule does hold due to a larger choice of the attentional gain control parameter. Thus the network experiences a sequence of recodings that ultimately stabilizes. In particular, on trial 4, node v_2 reads-out the subset template C, which mismatches the input pattern A. The numbers beneath the template symbols in row 4 describe the order of search. First, v_2 's template C mismatches A. Then v_1 's template B mismatches A. Finally A activates the uncommitted node v_3 , which resonates with F_1 as it learns the template A.

Scanning the rows of Figure 5b, we see that pattern A is coded by v_1 on trial 1; by v_3 on trials 4 and 6; and by v_4 on trial 9. On all future trials, input pattern A is coded by v_4 . Moreover, all the input patterns A, B, C, and D have learned a stable code by trial 9. Thus the code self-stabilizes by the second run through the input list ABCAD. On trials 11 through 15, and on all future trials, each input pattern chooses a different node ($A \rightarrow v_4$; $B \rightarrow v_1$; $C \rightarrow v_3$; $D \rightarrow v_2$). Each pattern belongs to a separate category because the vigilance parameter was chosen to be large in this example. Moreover, after code learning stabilizes, each input pattern directly activates its node in F_2 without undergoing any additional search. Thus after trial 9, only the "RES" symbol appears under the top-down templates. The patterns shown in any row between 9 and 15 provide a complete description of the learned code. Examples of how a novel exemplar can activate a previously learned category are found on trials 2 and 5 in Figures 5a and 5b. On trial 2, for example, pattern B is presented for the first time and directly accesses the category coded by v_1 , which was previously learned by pattern A on trial 1. In terminology from artificial intelligence, B activates the same categorical "pointer," or "marker," or "index" as in A. In so doing, B does not change the categorical "index," but it may change the categorical template, which determines which input patterns will also be coded by this index on future trials. The category does not change, but its invariants may change.

An example of how presentation of very different input patterns can influence the category of a fixed input pattern is found through consideration of trials 1, 4, and 9 in Figure 5b. These are the trials on which pattern A is recoded due to the intervening occurrence of other input patterns. On trial 1, pattern A is coded by v_1 . On trial 4, A is recoded by v_3 because pattern B has also been coded by v_1 and pattern C has been coded by v_2 in the interim. On trial 9, pattern A is recoded by v_4 both because pattern C has been recoded by v_3 and pattern D has been coded by v_2 in the interim.

In all of these transitions, the global structure of the input pattern determines which F_2 nodes will be activated, and global measures of pattern match at F_1 determine whether these nodes will be reset or allowed to resonate in STM.

8. Vigilance, Orienting, and Reset. We now show how matching within the attentional subsystem at F_1 determines whether or not the orienting subsystem will be activated, thereby leading to reset of the attentional subsystem at F_2 . The discussion can be broken into three parts:

A. Distinguishing Active Mismatch from Passive Inactivity

A severe mismatch at F_1 activates the orienting subsystem A. In the worst possible case of mismatch, none of the F_1 nodes can satisfy the 2/3 Rule, and thus no supraliminal activation of F_1 can occur. Thus in the worst case of mismatch, wherein F_1 becomes totally inactive, the orienting subsystem must surely be engaged.

On the other hand, F_1 may be inactive simply because no inputs whatsoever are being processed. In this case, activation of the orienting subsystem is not desired. How does the network compute the difference between active mismatch and passive inactivity at F_1 ?

This question led Grossberg [4] to assume that the bottom-up input source activates two parallel channels (Figure 2a). The attentional subsystem receives a specific input pattern at F_1 . The orienting subsystem receives convergent inputs at A from all the active

input pathways. Thus the orienting subsystem can be activated only when F_1 is actively processing bottom-up inputs.

B. Competition between the Attentional and Orienting Subsystems

How, then, is a bottom-up input prevented from resetting its own F_2 code? What mechanism prevents the activation of A by the bottom-up input from *always* resetting the STM representation at F_2 ? Clearly inhibitory pathways must exist from F_1 to A (Figure 2a). When F_1 is sufficiently active, it prevents the bottom-up input to A from generating a reset signal to F_2 . When activity at F_1 is attenuated due to mismatch, the orienting subsystem A is able to reset F_2 (Figure 2b,c,d). In this way, the orienting subsystem can distinguish between active mismatch and passive inactivity at F_1 .

Within this general framework, we now show how a finer analysis of network dynamics, with particular emphasis on the 2/3 Rule, leads to a vigilance mechanism capable of regulating how coarse the learned categories will be.

C. Collapse of Bottom-Up Activation due to Template Mismatch

Suppose that a bottom-up input pattern has activated F_1 and blocked activation of A (Figure 2a). Suppose, moreover, that F_1 activates an F_2 node which reads-out a template that badly mismatches the bottom-up input at F_1 (Figure 2b). Due to the 2/3 Rule, many of the F_1 nodes which were activated by the bottom-up input alone are suppressed by the top-down template. Suppose that this mismatch event causes a large collapse in the total activity across F_1 , and thus a large reduction in the total inhibition which F_1 delivers to A . If this reduction is sufficiently large, then the excitatory bottom-up input to A may succeed in generating a nonspecific reset signal from A to F_2 (Figure 2c).

In order to characterize when a reset signal will occur, we make the following natural assumptions. Suppose that an input pattern I sends positive signals to $|I|$ nodes of F_1 . Since every active input pathway projects to A , I generates a total input to A that is proportional to $|I|$. We suppose that A reacts linearly to the total input $\gamma |I|$. We also assume that each active F_1 node generates an inhibitory signal of fixed size to A . Since every active F_1 node projects to A , the total inhibitory input $\delta |X|$ from F_1 to A is proportional to the number $|X|$ of active F_1 nodes. When $\gamma |I| > \delta |X|$, A receives a net excitatory signal and generates a nonspecific reset signal to F_2 (Figure 2c).

In response to a bottom-up input pattern I of size $|I|$, as in Figure 2a, the total inhibitory input from F_1 to A equals $\delta |I|$, so the net input to A equals $(\gamma - \delta) |I|$. In order to prevent A from firing in this case (Figure 2a), we assume that $\delta \geq \gamma$. We call

$$\rho = \frac{\gamma}{\delta} \quad (6)$$

the *vigilance parameter* of the orienting subsystem. The constraints $\delta \geq \gamma \geq 0$ are equivalent to $0 \leq \rho \leq 1$. The size of ρ determines the proportion of the input pattern which must be matched in order to prevent reset.

When both a bottom-up input I and a top-down template $V^{(j)}$ are simultaneously active (Figure 2b), the 2/3 Rule implies that the total inhibitory signal from F_1 to A equals $\delta |V^{(j)} \cap I|$. In this case, the orienting subsystem is activated only if

$$\gamma |I| > \delta |V^{(j)} \cap I|; \quad (7)$$

that is, if

$$\frac{|V^{(j)} \cap I|}{|I|} < \rho. \quad (8)$$

In order to illustrate how the network codifies a series of patterns, we show in Figure 6 the first 20 trials of a simulation using alphabet letters as input patterns. In Figure 6a,

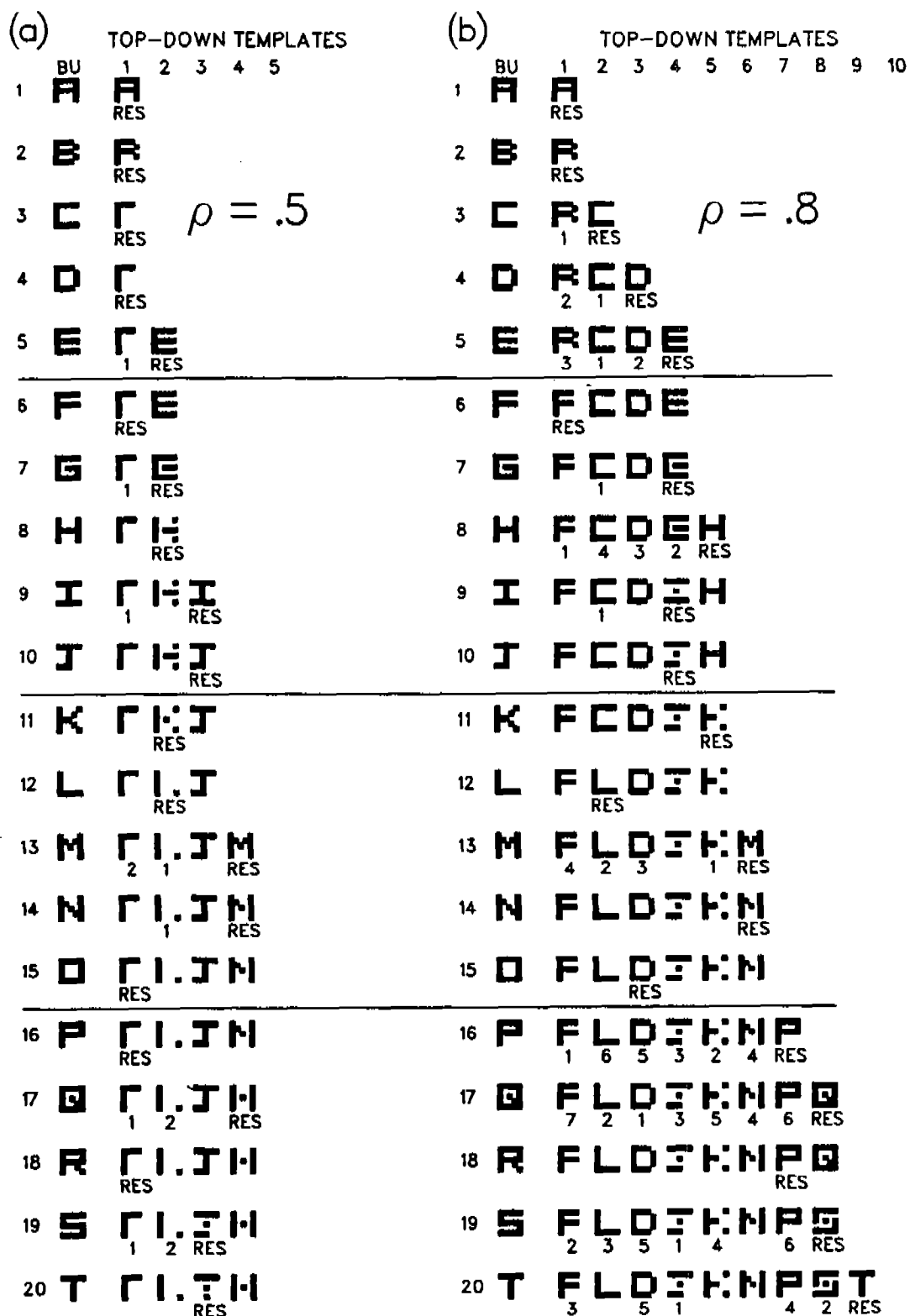


Figure 6. Alphabet learning: Different vigilance levels cause different numbers of letter categories to form.

the vigilance parameter $\rho = .5$. In Figure 6b, $\rho = .8$. Three properties are notable in these simulations. First, choosing a different vigilance parameter can determine different coding histories, such that higher vigilance induces coding into finer categories. Second, the network modifies its search order on each trial to reflect the cumulative effects of prior learning, and bypasses the orienting system to directly access categories after learning has taken place. Third, the templates of coarser categories tend to be more abstract because they must approximately match a larger number of input pattern exemplars.

Given $\rho = .5$, the network groups the 26 letter patterns into 8 stable categories within 3 presentations. In this simulation, F_2 contains 15 nodes. Thus 7 nodes remain uncoded because the network self-stabilizes its learning after satisfying criteria of vigilance and global code self-consistency. Given $\rho = .8$ and 15 F_2 nodes, the network groups 25 of the 26 letters into 15 stable categories within 3 presentations. The 26th letter is rejected by the network in order to self-stabilize its learning while satisfying its criteria of vigilance and global code self-consistency. These simulations show that the network's use of processing resources depends upon an evolving dynamical organization with globally context-sensitive properties. This class of networks is capable of organizing arbitrary sequences of arbitrarily complex input patterns into stable categories subject to the constraints of vigilance, global code self-consistency, and number of nodes in F_1 and F_2 .

APPENDIX NETWORK EQUATIONS

STM Equations

The STM activity of any node v_k in F_1 or F_2 obeys a membrane equation of the form

$$\frac{d}{dt}x_k = -Ax_k + (B - Cx_k)J_k^+ - Dx_kJ_k^-, \quad (A1)$$

where J_k^+ and J_k^- are the total excitatory input and total inhibitory input, respectively, to v_k and A, B, C, D are nonnegative parameters. If $C > 0$, then the STM activity $x_k(t)$ remains within the finite interval $[0, BC^{-1}]$ no matter how large the inputs J_k^+ and J_k^- are chosen.

We denote nodes in F_1 by v_i , where $i = 1, 2, \dots, M$. We denote nodes in F_2 by v_j , where $j = M + 1, M + 2, \dots, N$. Thus by (A1),

$$\frac{d}{dt}x_i = -A_1x_i + (B_1 - C_1x_i)J_i^+ - D_1x_iJ_i^- \quad (A2)$$

and

$$\frac{d}{dt}x_j = -A_2x_j + (B_2 - C_2x_j)J_j^+ - D_2x_jJ_j^-. \quad (A3)$$

The input J_i^+ is a sum of the bottom-up input I_i and the top-down template

$$V_i = \sum_j f(x_j)z_{ji}, \quad (A4)$$

that is,

$$J_i^+ = I_i + V_i. \quad (A5)$$

where $f(x_j)$ is the signal generated by activity x_j of v_j , and z_{ji} is the LTM trace in the pathway from v_j to v_i .

The inhibitory input J_i^- controls the attentional gain:

$$J_i^- = F \sum_j f(x_j). \quad (A6)$$

Thus $J_i^- = 0$ if and only if F_2 is inactive (Figure 4).

The inputs and parameters of STM activities in F_2 were chosen so that the F_2 node which received the largest input from F_1 wins the competition for STM activity. Theorems show how these parameters can be chosen [15–17]. The inputs J_j^+ and J_j^- have the following form.

Input J_j^+ adds a positive feedback signal $g(x_j)$ from v_j to itself to the bottom-up adaptive filter input

$$T_j = \sum_i h(x_i) z_{ij}. \quad (A7)$$

that is,

$$J_j^+ = g(x_j) + T_j. \quad (A8)$$

where $h(x_i)$ is the signal emitted by v_i and z_{ij} is the LTM trace in the pathway from v_i to v_j . Input J_j^- adds up negative feedback signals $g(x_k)$ from all the other nodes in F_2 :

$$J_j^- = \sum_{k \neq j} g(x_k). \quad (A9)$$

Such a network behaves approximately like a binary switching circuit:

$$x_j = \begin{cases} G & \text{if } T_j > \max(T_k : k \neq j) \\ 0 & \text{otherwise.} \end{cases} \quad (A10)$$

LTM Equations

The LTM trace of the bottom-up pathway from v_i to v_j obeys a learning equation of the form

$$\frac{d}{dt} z_{ij} = f(x_j) [-H_{ij} z_{ij} + K h(x_i)]. \quad (A11)$$

In (A11), term $f(x_j)$ is a postsynaptic sampling, or learning, signal because $f(x_j) = 0$ implies $\frac{d}{dt} z_{ij} = 0$. Term $f(x_j)$ is also the output signal of v_j to pathways from v_j to F_1 , as in (A4).

The LTM trace of the top-down pathway from v_j to v_i also obeys a learning equation of the form

$$\frac{d}{dt} z_{ji} = f(x_j) [-H_{ji} z_{ji} + K h(x_i)]. \quad (A12)$$

In the present simulations, the simplest choice of H_{ji} was made for the top-down LTM traces:

$$H_{ji} = H = \text{constant.} \quad (A13)$$

A more complex choice of H_{ji} was made for the bottom-up LTM traces. This was done to directly generate the Weber Law Rule [2] via the bottom-up LTM process itself. The Weber Law Rule can also be generated indirectly by exploiting a Weber Law property of competitive STM interactions across F_1 . Such an indirect instantiation of the Weber Law Rule enjoys several advantages. In particular, it would enable us to also choose $H_{ji} = H = \text{constant}$. Instead, we allowed the bottom-up LTM traces at each node v_j to compete among themselves for synaptic sites. Malsburg and Willshaw [18] have used a related idea in their model of retinotectal development. In the present usage, it was essential to choose a shunting competition to generate the Weber Law Rule, unlike the Malsburg and Willshaw usage. Thus we let

$$H_{ji} = Lh(x_i) + \sum_{k \neq i} h(x_k). \quad (\text{A14})$$

A physical interpretation of this choice can be seen by rewriting (A11) in the form

$$\frac{d}{dt} z_{ij} = f(x_j) [(K - Lz_{ij})h(x_i) - z_{ij} \sum_{k \neq i} h(x_k)]. \quad (\text{A15})$$

By (A15), when the postsynaptic signal $f(x_j)$ is positive, a positive presynaptic signal $h(x_i)$ commits receptor sites to the LTM process z_{ij} at a rate $(K - Lz_{ij})h(x_i)f(x_j)$. Simultaneously, signals $h(x_k)$, $k \neq i$, which reach v_j at different regions of the v_j membrane compete for sites which are already committed to z_{ij} via the mass action competitive terms $-z_{ij}f(x_j)h(x_k)$. When z_{ij} equilibrates to these competing signals,

$$z_{ij} = \frac{Kh(x_i)}{(L-1)h(x_i) + \sum_k h(x_k)}. \quad (\text{A16})$$

The signal function $h(w)$ was chosen to rise quickly from 0 to 1 at a threshold activity level w_0 . Thus if v_i is a suprathreshold node in F_1 , (A16) approximates

$$z_{ij} \cong \frac{K}{(L-1) + |X|} \quad (\text{A17})$$

where $|X|$ is the number of active nodes in F_1 . Term z_{ij} obeys a Weber Law Rule if $L > 1$.

STM Reset System

The simplest possible mismatch-mediated activation of A and STM reset of F_2 by A were implemented in the simulations. As outlined in Section 3, each active input pathway sends an excitatory signal of size γ to A . Potentials x_i of F_1 which exceed a signal threshold T generate an inhibitory signal of size $-\delta$ to A . Population A , in turn, generates a nonspecific reset wave to F_2 whenever

$$\gamma |I| - \delta |X| > 0. \quad (\text{A18})$$

where I is the current input pattern and $|X|$ is the number of nodes across F_1 such that $x_i > T$. The nonspecific reset wave shuts off the active F_2 node until the input pattern I

shuts off. Thus (A10) must be modified to shut off all F_2 nodes which have been reset by A during the presentation of I .

REFERENCES

- [1] Basar, E., Flohr, H., Haken, H., and Mandell, A.J. (Eds.), **Synergetics of the brain**. New York: Springer-Verlag, 1983.
- [2] Carpenter, G.A. and Grossberg, S., Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. In J. Davis, R. Newburgh, and E. Wegman (Eds.), **Brain structure, learning, and memory**. AAAS Symposium Series, 1985.
- [3] Carpenter, G.A. and Grossberg, S., Self-organization of neural recognition categories. In preparation. 1985.
- [4] Grossberg, S., How does a brain build a cognitive code? *Psychological Review*, 1980, **87**, 1-51.
- [5] Grossberg, S., A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology**, Vol. 5. New York: Academic Press, 1978, pp.233-374.
- [6] Grossberg, S., The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E.C. Schwab and H.C. Nusbaum (Eds.), **Perception of speech and visual form: Theoretical issues, models, and research**. New York: Academic Press, 1985.
- [7] Grossberg, S. and Stone, G.O., Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review*, in press, 1985.
- [8] Grossberg, S. and Mingolla, E., Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review*, 1985, **92**, 173-211.
- [9] Grossberg, S. and Mingolla, E., Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Perception and Psychophysics*, in press, 1985.
- [10] Cohen, M.A. and Grossberg, S., Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, in press. 1985.
- [11] Grossberg, S., Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 1976, **23**, 121-134.
- [12] Grossberg, S., Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, 1976, **23**, 187-202.
- [13] Grossberg, S., Some psychophysiological and pharmacological correlates of a developmental, cognitive, and motivational theory. In R. Karrer, J. Cohen, and P. Tueting (Eds.), **Brain and information: Event related potentials**. New York: New York Academy of Sciences, 1984.
- [14] Grossberg, S., The quantized geometry of visual space: The coherent computation of depth, form, and lightness. *Behavioral and Brain Sciences*, 1983, **6**, 625-692.
- [15] Elias, S.A. and Grossberg, S., Pattern formation, contrast control, and oscillations in the short term memory of shunting on-center off-surround networks. *Biological Cybernetics*, 1975, **20**, 69-98.
- [16] Grossberg, S., Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 1973, **52**, 217-257.
- [17] Grossberg, S. and Levine, D.S., Some developmental and attentional biases in the contrast enhancement and short term memory of recurrent neural networks. *Journal of Theoretical Biology*, 1975, **53**, 341-380.

- [18] Malsburg, C. von der and Willshaw, D.J., Differential equations for the development of topological nerve fibre projections. In S.Grossberg (Ed.), **Mathematical psychology and psychophysiology**. Providence, RI: American Mathematical Society, 1981.

NONLINEAR NEURAL DYNAMICS OF VISUAL SEGMENTATION

Stephen Grossberg[†]
and
Ennio Mingolla[‡]
Center for Adaptive Systems
Boston University
111 Cummington Street
Boston, Massachusetts 02215

ABSTRACT. Neural network models of boundary completion, textural segmentation, and perceptual grouping possess novel mathematical properties with implications in psychology, neurobiology, artificial intelligence, geometry, statistical mechanics, and decision theory. These neural circuits explain color and boundary data as well as textural grouping phenomena. Some of the circuits have recently received experimental support from neurophysiological recordings from monkey visual cortex. The circuits suggest a new approach to the design of computer vision systems.

1. Nonlinear Dynamical Systems in Visual Perception, Neurobiology, and Artificial Intelligence. The venerable subject of visual perception offers the modern student of nonlinear dynamical systems a wealth of new phenomena and concepts that are ripe for development. This is partly because the profound issues of visual perception, like those of number theory, are often revealed through the use of immediately accessible materials. Just as many of the deepest questions about the integers can be presented to any student, many of the deepest phenomena concerning visual perception can be readily seen by a casual observer (Figure 1). Easily stated number theoretic questions have led to some of the world's most profound mathematics. Efforts to explain visual phenomena have also given rise to increasingly abstract concepts since the pioneering work of Helmholtz, Mach, and Maxwell.

The perceptual processing theory that we and our colleagues are now developing has led to the discovery of new dynamical systems models of nonlinear competition, cooperation, diffusion, and resonance. These models have been used to quantitatively simulate on the computer many visual percepts that have not been explained by previous theories, to link these percepts to recent neurophysiological and anatomical data (Cohen and Grossberg, 1984a, 1984b; Grossberg, 1983a, 1983b, 1984; Grossberg and Mingolla, 1985a, 1985b), and to make some neural predictions which have recently been confirmed (Grossberg and Mingolla, 1985a).

As befits formal theories which model important physical phenomena, these dynamical systems exhibit a wide variety of mathematical properties and have led to new types of theorems. For example, certain dynamical systems of this type have the remarkable

[†] Supported in part by the Air Force Office of Scientific Research (AFOSR 85-0149) and the Army Research Office (ARO DAAG-29-85-K-0095).

[‡] Supported in part by the Air Force Office of Scientific Research (AFOSR 85-0149).

Acknowledgments: We wish to thank Cynthia Suchta for her valuable assistance in the preparation of the manuscript and illustrations.

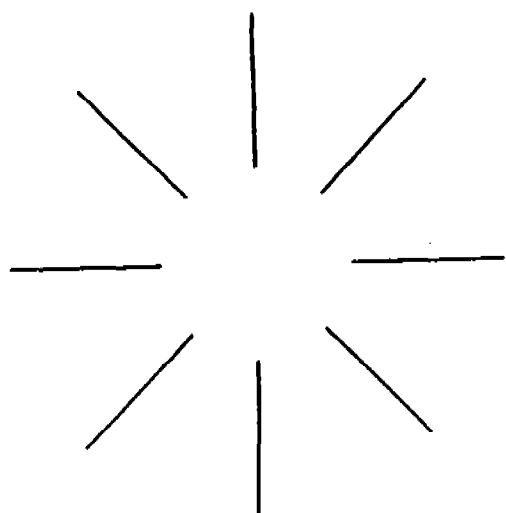
property that, despite their very high dimension, they always approach one of perhaps infinitely many equilibrium points given any physical choice of initial values and parameters (Cohen and Grossberg, 1983; Grossberg, 1978, 1980). A wide variety of complex oscillations, including period doubling and chaotic oscillations, are known to be solutions of closely related nonlinear dynamical systems (Carpenter and Grossberg, 1983, 1984, 1985; Cohen and Grossberg, 1983; Grossberg, 1980). Although a great deal is now known about these several types of dynamical behavior, it remains to completely characterize the mathematical conditions under which they obtain. In addition, such dynamical systems describe a new type of dynamical geometry. Familiar geometrical objects such as curves and surfaces are shown to be characterized by neural processes in a manner that is radically different from the axioms of geometry. These systems also illustrate that the brain realizes novel principles of nonequilibrium statistical mechanics and nonstationary decision theory. New circuit designs for a sophisticated type of computer vision are also suggested. Thus the present theory defines a mathematical framework in which new mathematical ideas about perception, neurobiology, dynamical systems, geometry, statistical mechanics, decision theory, and artificial intelligence are developing side-by-side.

This theory has emerged from an analysis of how brain designs achieve informative visual representations of the external world that are much more veridical than the retinal sensory data from which they are derived. Of special interest are issues concerning how distributed patterns of locally ambiguous visual cues can be used to generate unambiguous global percepts through the mediation of nonlinear dynamical interactions. The present article describes some of the issues that led to the theory, introduces the dynamical systems themselves, and illustrates some of their behaviors using computer simulations.

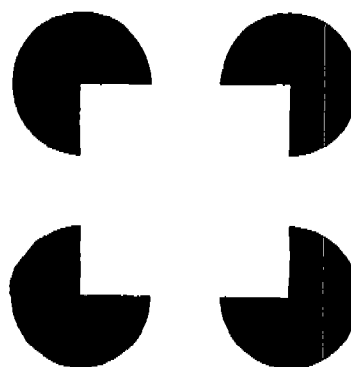
2. From Noisy Retina to Coherent Percept. The limitations of sensory information on the retina present a formidable challenge to any perceptual theory. For example, light passes through a thicket of retinal veins before it reaches retinal photoreceptors. The percepts of human observers are fortunately not distorted by their retinal veins during normal vision. The retinal veins are not perceived because of the action of mechanisms which attenuate the perception of images that are stabilized with respect to the retina. Suppressing the percept of the stabilized veins is, however, far from sufficient to generate an unambiguous percept. This is because the images that reach the retina are often occluded and segmented by the veins in several places. Even a single edge can be broken into several disjoint components. Somehow in the final percept broken retinal contours are completed, and occluded retinal color and brightness signals are filled-in. These completed and filled-in percepts are, in a strict mechanistic sense, illusory percepts.

Observers are not aware of which parts of a perceived edge are "real" and which are "illusory." Thus it is not surprising that illusory percepts have provided important clues for deriving the present theory, and that the theory can be used to analyse many percepts of illusory boundaries, colors, and brightnesses. Of special interest are experiments of Krauskopf (1963) and Yarbus (1967). These experiments show that, if certain scenic edges are artificially stabilized with respect to the retina, then colors and brightnesses which were previously bounded by these edges can flow across, or fill-in, the percept until they are contained by the next perceptually significant boundary (Figure 1c). Such results clarify how the visual system synthesized boundaries, both "real" and "illusory," that it selects as perceptually significant, and how featural filling-in occurs within these boundaries.

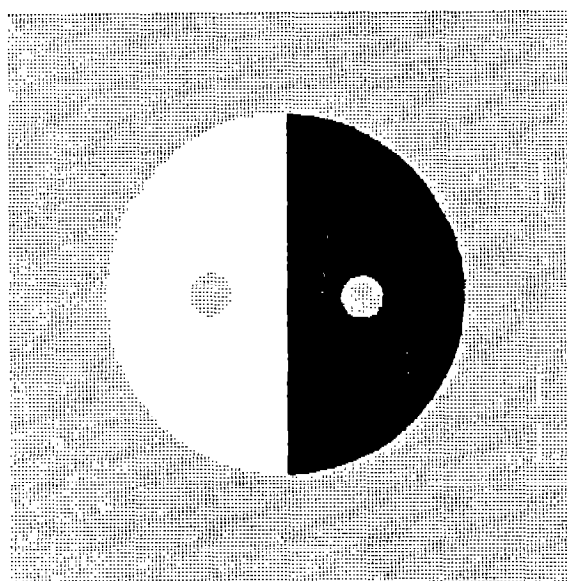
The retinal veins are not the only imperfection in the retinal image. The visual world is typically viewed in inhomogeneous lighting conditions. The scenic luminances that reach the retina thus confound fluctuating lighting conditions with invariant object colors and brightnesses. It has been known since the time of Helmholtz (1962) that the brain somehow discounts the spurious illuminants to generate color and brightness percepts that are more veridical than those in the retinal image. Land (1977) has shown that the perceived colors within a picture constructed from overlapping patches of color are determined by



(a)



(b)



(c)

Figure 1: Some perceptual illusions. (a) A bright illusory disc is induced perpendicular to the ends of the radial lines (Kennedy, 1979). (b) A bright illusory square is induced parallel to four black pac-man figures (Kanizsa, 1974). (c) When the edges of the large circle and the vertical line are stabilized on the retina, the red color (dots) outside the large circle envelopes the black and white hemi-discs except within the small red circles whose edges are not stabilized (Yarbus, 1967). The red inside the circles looks brighter (right) or darker (left) than the enveloping red.

the relative contrasts of the edges between successive patches. The luminances within the

interiors of each patch are somehow discounted.

This type of result also points to the action of a filling-in process. Were it not possible to fill-in among non-discounted edges, we would perceive a world of boundaries or cartoons. Since edges are used to generate the final filled-in percept, a good theory must define edge computations in a way that can achieve this goal.

To explain how this is accomplished, the present theory hypothesizes that at least two fundamentally different types of edge processing occur in parallel during human visual perception. As in the Land color demonstrations, edge computations in the present theory are used to generate a final percept, but edge computations by themselves do not constitute the final percept. These operations have been used to provide a physical interpretation and generalization (Grossberg, 1984) of the Land retinex theory of color and brightness perception (Land, 1977).

3. The Boundary Contour System and the Feature Contour System. The theory asserts that two distinct types of edge, or contour, computations are carried out within two parallel systems, which we call the Boundary Contour System and the Feature Contour System (Figure 2). Boundary Contour signals are used to synthesize the boundaries that the visual system selects as perceptually significant. Feature Contour signals initiate the filling-in processes whereby brightnesses and colors spread until they hit their first boundary contours, or are attenuated due to their spatial spread if no boundary contours intervene. The Boundary Contour System is defined by a nonlinear cooperative-competitive dynamical system whose inputs are derived from a nonlinear filter. The Feature Contour System is defined by a nonlinear diffusion process whose inputs are derived from a nonlinear filter and whose diffusion coefficients are controlled by the Boundary Contour System.

These two types of contour systems have not been identified in previous perceptual theories. One reason for this delay is that each scenic edge can activate both systems. Only the net effect of the interaction between systems is perceived. Another reason is that boundary contours are not, by themselves, visible. They become visible by restricting the filling-in that is triggered by feature contour signals.

If boundary contours are, by themselves, invisible, then how can we be sure that they exist? This is accomplished by an analysis of perceptual data which, by inference, reveals the *different rules* whereby these contours are computed. The main rules are the following ones.

4. Boundary Contours and Boundary Completion. The process whereby boundary contours are built up is initiated by the activation of oriented masks, or elongated receptive fields, at each position of perceptual space (Hubel and Wiesel, 1977).

A. Orientation and Contrast: The output signal that is generated by an oriented mask to the next processing stage is sensitive to the *orientation* and to the *amount* of contrast, but not to the *direction* of contrast, at an edge of a scene. Thus a vertical boundary contour can be activated by either a close-to-vertical light-dark edge or a close-to-vertical dark-light edge at a fixed scenic position.

B. Short-Range Oriented Shunting Competition: A mask of fixed orientation excites the like-oriented cells activated at its location and inhibits the like-oriented cells activated at nearby locations (stage w of Figure 3a). In other words, an on-center off-surround organization of like-oriented cell interactions exists around each perceptual location. The off-surround inhibition is of shunting, or divisive, type, rather than being subtractive, to prevent a straight scenic edge of fixed contrast from self-annihilating its own percept.

C. Oriented Tonic Opponent Processing: At each perceptual location, the cells that react to perpendicularly oriented masks compete (stage x in Figure 3a). This competition

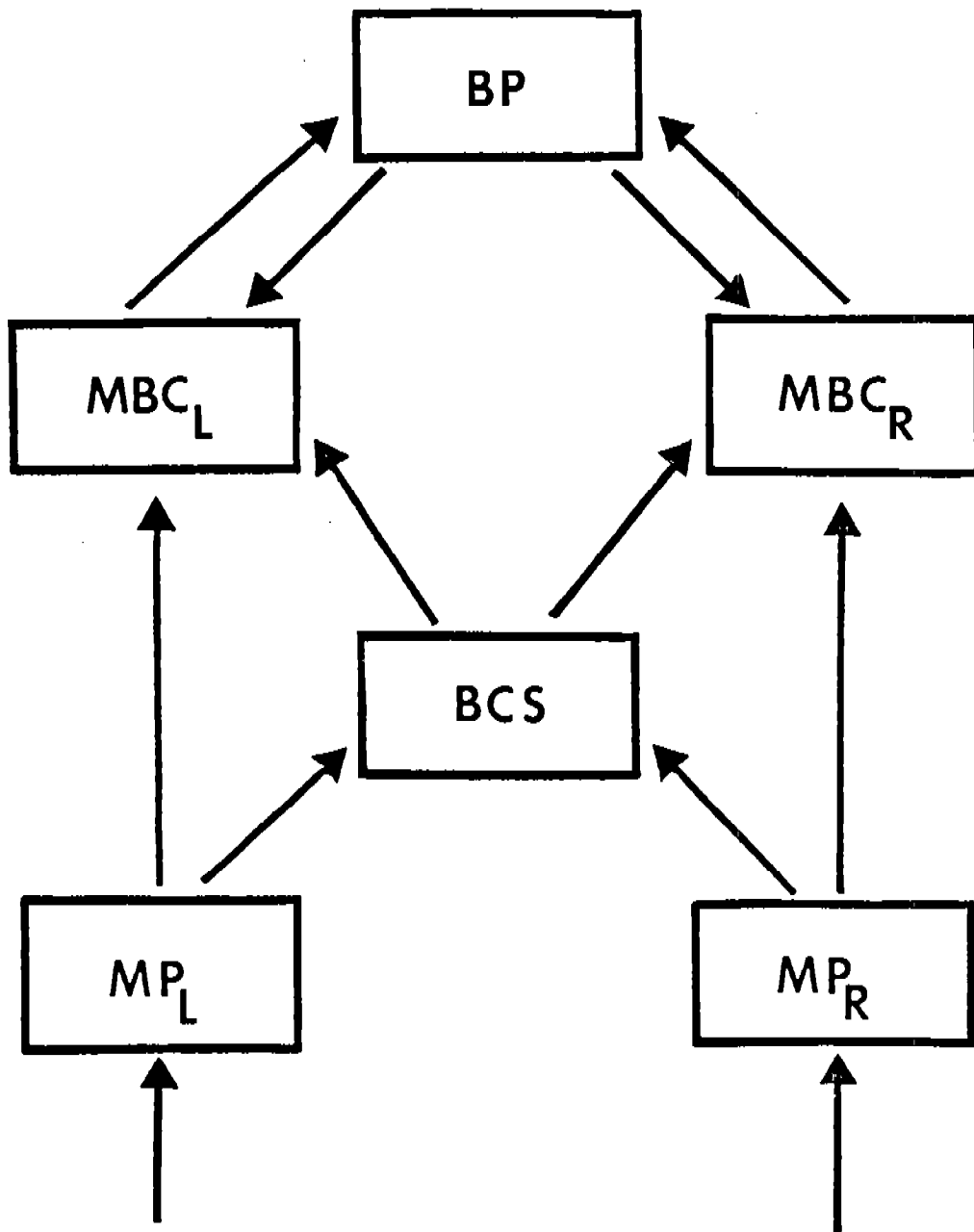


Figure 2: A macrocircuit of processing stages: The functional names of the abbreviated stages are: MP_L = Left Monocular Preprocessing stage; MP_R = Right Monocular Preprocessing stage; BCS = Boundary Contour Synthesis stage; MBC_L = Left-Monocular Brightness and Color stage; MBC_R = Right-Monocular Brightness and Color stage; and BP = Binocular Percept stage. Neural interpretations of the abbreviated stages have been given elsewhere (Grossberg, 1984).

defines a push-pull opponent process: If a given orientation is inhibited, then its perpendicular orientation is disinhibited. Hence the opponent processes are tonically active.

D. Normalization and Ratio Scale: A final stage of orientational competition at each position tends to normalize, or adapt, the total output of the boundary signals corresponding to that position (stage y of Figure 3a). The orientationally tuned suprathreshold outputs corresponding to each position of perceptual space thus tend to be ratios of a conserved total activity, such that orthogonal orientations cannot simultaneously possess

positive ratios.

E. *Long-Range Oriented Cooperation and Boundary Completion*: Like-oriented cells which survive the competitive interactions (A)–(D), and which have approximately aligned orientations across perceptual space, can activate an oriented cooperative process z (Figure 3a). The cooperative process, in turn, feeds back to the competitive process via the chain of processes $z \rightarrow v \rightarrow w$. The feedback exchange between these competitive and cooperative interactions is capable of rapidly synthesizing sharp boundaries, both real and illusory, and of rapidly segmenting a scene. The dynamical system which defines the Boundary Contour System is now described.

Indices (i, j) denote position in a two-dimensional lattice and k denotes an orientation. Let S_{ij} equal the input to position (i, j) . Divide the oriented mask with position (i, j) and orientation k into a left-half L_{ijk} and a right-half R_{ijk} . Then

$$J_{ijk} = \frac{[U_{ijk} - \alpha V_{ijk}]^+ + [V_{ijk} - \alpha U_{ijk}]^+}{1 + \beta(U_{ijk} + V_{ijk})} \quad (1)$$

where

$$U_{ijk} = \sum_{(p,q) \in L_{ijk}} S_{pq}, \quad (2)$$

$$V_{ijk} = \sum_{(p,q) \in R_{ijk}} S_{pq}, \quad (3)$$

and the notation $[p]^+ = \max(p, 0)$;

$$\frac{d}{dt} v_{ijk} = -v_{ijk} + f(z_{ijk}) - v_{ijk} \sum_{(p,q)} f(z_{pqk}) A_{pqij}, \quad (4)$$

where

$$A_{pqij} = \begin{cases} A & \text{if } (p-i)^2 + (q-j)^2 \leq A_0 \\ 0 & \text{otherwise;} \end{cases} \quad (5)$$

$$\frac{d}{dt} w_{ijk} = -w_{ijk} + I + J_{ijk} + v_{ijk} - w_{ijk} \sum_{(p,q)} J_{pqk} B_{pqij}, \quad (6)$$

where

$$B_{pqij} = \begin{cases} B & \text{if } (p-i)^2 + (q-j)^2 \leq B_0 \\ 0 & \text{otherwise;} \end{cases} \quad (7)$$

$$\frac{d}{dt} x_{ijk} = -x_{ijk} + C[w_{ijk} - w_{ijK}]^+, \quad (8)$$

where K is perpendicular to k ;

$$\frac{d}{dt} y_{ijk} = -Dy_{ijk} + (E - y_{ijk})x_{ijk} - y_{ijk} \sum_{m \neq k} x_{ijm}; \quad (9)$$

$$\frac{d}{dt} z_{ijk} = -z_{ijk} + g\left(\sum_{(p,q,r)} [y_{pqr} - y_{pqR}] F_{pqij}^{(r,k)}\right) + g\left(\sum_{(p,q,r)} [y_{pqr} - y_{pqR}] G_{pqij}^{(r,k)}\right), \quad (10)$$

where R is perpendicular to r , and the cooperative kernels $F_{pq}^{(r,k)}$ and $G_{pq}^{(r,k)}$ are defined by

$$F_{pqij}^{(r,k)} = \left[\exp \left[-2 \left(\frac{L_{pqij}}{M} - 1 \right)^2 \right] \left| \cos(N_{pqij} - r) \right| \right]^P \left[\cos(N_{pqij} - k) \right]^Q \quad (11)$$

and

$$G_{pqij}^{(r,k)} = \left[-\exp \left[-2 \left(\frac{L_{pqij}}{M} - 1 \right)^2 \right] \left| \cos(N_{pqij} - r) \right| \right]^P \left[\cos(N_{pqij} - k) \right]^Q, \quad (12)$$

where

$$L_{pqij} = \sqrt{(p-i)^2 + (q-j)^2}, \quad (13)$$

and

$$N_{pqij} = \arctan \left(\frac{q-j}{p-i} \right). \quad (14)$$

The nonlinear signal functions f in (4) and g in (10) are of the form

$$f(\xi) = [\xi - R]^+ \quad (15)$$

and

$$g(\xi) = \frac{[\xi]^+}{T + [\xi]^+}. \quad (16)$$

Figure 1b illustrates the relevance of some of these rules to perceptual phenomena. In Figure 1b, an illusory square can be seen. The theory suggests that the vertical illusory boundary contours are completed by the cooperative process, described in Section 4E, in response to the vertically oriented masks, described in Section 4A. An illusory square can also be seen if the two black pac-man figures at the bottom of Figure 1b are replaced by white pac-man figures, and the white background is replaced by a grey background. The black pac-man figures form a dark-light edge with respect to the grey background. The white pac-man figures form light-dark edges with respect to the grey background. The visibility of vertical illusory boundaries shows that a process exists that is capable of completing boundaries between edges with opposite directions of contrast. The boundary completion process is thus sensitive to orientational alignment across perceptual space, as in Section 4E, and to amount of contrast but not to direction of contrast, as in Section 4A.

5. Feature Contours and Diffusive Filling-In. The Feature Contour process obeys different rules of contrast than those governing the Boundary Contour process.

A. Contrast: The feature contour process is sensitive to *direction* of contrast as well as to *amount* of contrast, unlike the boundary contour process. Thus to compute the relative brightness across a scenic boundary, one needs to keep track of which side of the scenic boundary has a larger reflectance. Direction of contrast is also used to determine which side of a red-green scenic boundary is red and which is green. The different sensitivities of the two contour systems to direction of contrast is one of their most important dissociating properties.

The Feature Contour process also obeys different rules of spatial interactions than those governing the Boundary Contour process.

B. Diffusive Filling-In: Boundary contours activate a boundary completion process that synthesizes the boundaries which define monocular perceptual domains. Feature contours activate a diffusive filling-in process that spreads featural qualities, such as brightness or

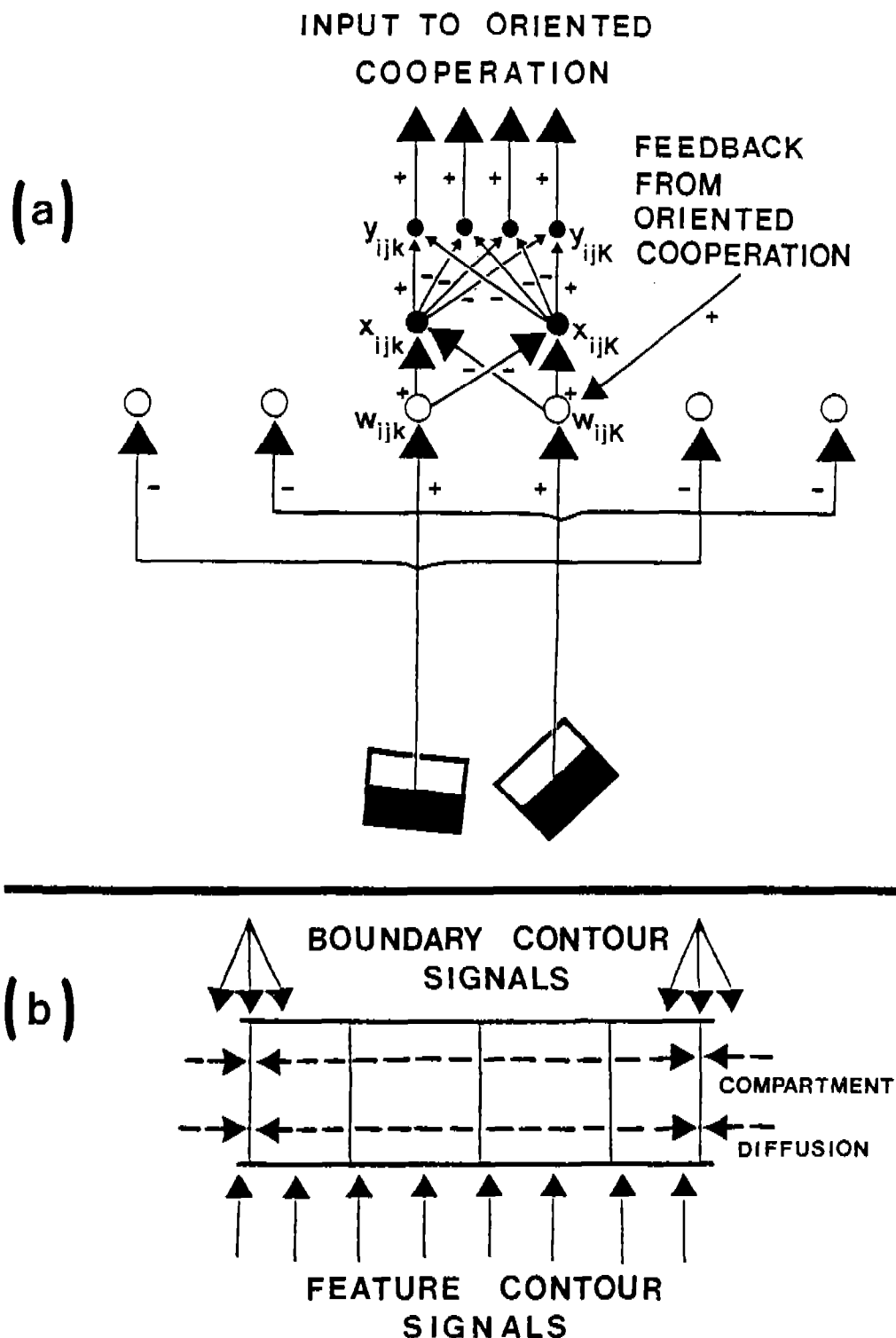


Figure 3: (a) Boundary Contour completion: An on-center off-surround interaction between like-oriented cells representing nearby positions (stage w) is followed by tonic opponent processing between orthogonally oriented cells at each position (stage x). Then the total suprathreshold activity at each position is normalized (stage y) before each orientation feeds into long-range cooperation (stage z) among similarly oriented and spatially aligned cells. Positive feedback from the cooperative process feeds back to like-oriented cells at the opponent processing stage. (b) Feature Contour filling-in: Feature Contour signals activate cell compartments which permit rapid lateral diffusion of electrical potential across their membranes, except at those membranes which receive Boundary Contour signals from the BCS stage of Figure 2.

color, across these perceptual domains, as in the Yarbus (1967) demonstration (Figure 1c). Figure 3b depicts the main properties of this filling-in process.

It is assumed that featural filling-in occurs within a array of closely coupled cell compartments. A feature contour input signal to a cell of the array triggers a rapid diffusion of electrical potential across the compartment membranes of neighboring cells. This diffusion spreads with a space constant that depends upon the electrical properties of both the cell interiors and their membranes.

A Boundary Contour signal is assumed to decrease the diffusion constant of its target cell membranes within the cell syncytium. It does so by acting as an inhibitory gating signal that causes an increase in membrane resistance. At the same time that a boundary contour signal attenuates the filling-in process at its target cells, it acts to inhibit the potentials of these cells. The nonlinear diffusion process which instantiates properties (5A) and (5B) is rigorously defined in Cohen and Grossberg (1984b).

6. Boundary-Feature Trade-Off: New Axioms of Geometry. The theory's rules seem natural when one realizes that the rules of each contour system are designed to offset insufficiencies of the other contour system. This realization also leads to the conclusion that neural representations of geometrical objects do not obey the axioms of geometry. The Boundary Contour system, by itself, could at best generate a world of boundaries or cartoons. The Feature Contour system, by itself, could at best generate a world of formless qualities. Once one accepts that featural filling-in spreads over perceptually ambiguous regions until reaching a boundary contour, it becomes an urgent task to synthesize boundaries capable of containing the featural flow. Orientationally tuned input masks, or receptive fields, are needed to initiate the process of building up these boundary contours. Orientationally tuned input masks are, however, insensitive to orientation at the ends of lines and at object corners. This breakdown is illustrated by the simulation in Figure 4a, which depicts the reaction of a lattice of orientationally tuned masks to a thin vertical line (Grossberg and Mingolla, 1985a). Without further processing of the mask outputs, featural quality could flow freely out of every line end or corner. Such a flow does, in fact, occasionally occur *in vivo* in response to certain scenes, as in neon color spreading (Redies and Spillmann, 1981; van Tuijl, 1975), or in the featural flow that occurs over retinally stabilized scenic edges (Krauskopf, 1963; Yarbus, 1967).

To offset this difficulty under normal circumstances, we suggest that the boundary system initiates the several stages of competitive interaction described in Section 4 to compensate for orientational insensitivity at line ends and at corners. Figure 4 shows how these competitive interactions generate horizontal boundary signals at the end of the thin vertical line that help to prevent the flow of featural quality out of the line. Such boundary signals are said to be generated by *end cutting*, or *orthogonal induction*. Thus every line end is an illusory percept that is reconstructed at a high level of visual processing.

The output pattern of the competitive process triggers, in its turn, a long-range oriented cooperative process, which builds up the completed boundary contours. Computer simulations (Figure 5) show that this process is capable of quickly building sharp boundaries that span widely separated input masks while suppressing spurious noise. Figure 5 illustrates that the process which synthesizes a percept of line may not even form a connected set until the system approaches equilibrium. Thus a perceived line is, in part, an equilibrium solution of a mixed cooperative-competitive nonlinear feedback network, rather than a connected set of points.

The illusory circle in Figure 1a can now be analysed as a result of orthogonal end cutting followed by oriented boundary cooperation. Many previously mysterious percepts have been analysed as consequences of the fundamental Boundary-Feature Trade-Off. All of these analyses suggest that the way in which we perceive geometrical objects does not correspond well to the classical axioms of geometry.

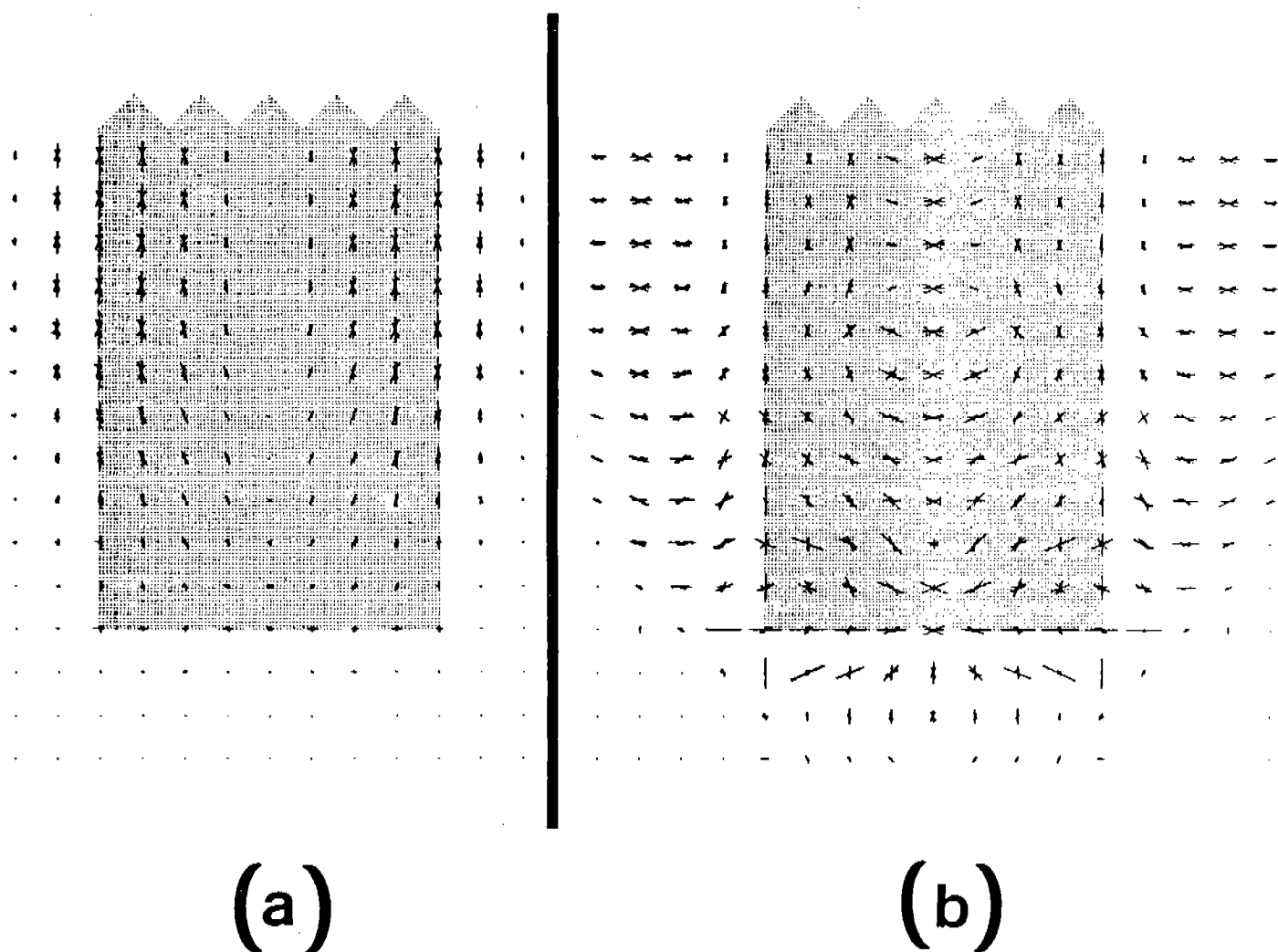


Figure 4: Boundary-Feature Trade-Off: (a) Response of an orientation field to a thin vertical line which looks blown up relative to receptive field sizes. Lengths and orientations of lines encode relative amounts of activation and orientations of the oriented masks at the corresponding bar positions. Orientational tuning breaks down at the end of the bar. (b) End-cutting generates “illusory” horizontal activations at the end of the line in response to the orientation field in (a), to offset the breakdown in orientational tuning. The end-cutting pattern is the output of stage γ of Figure 3a. Thus “every line end is illusory,” in sharp contrast to the axioms of geometry.

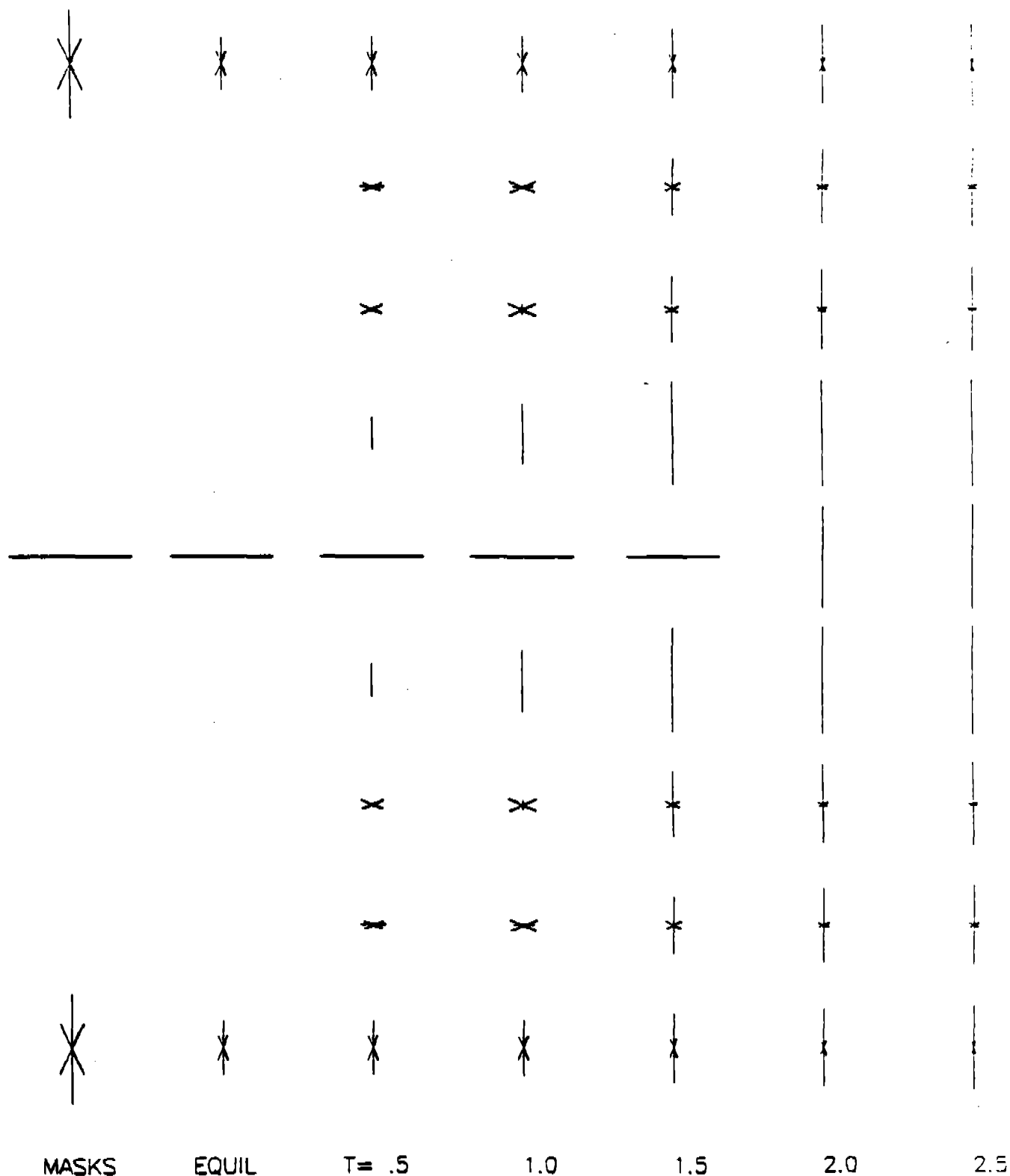


Figure 5: Each column depicts a different time during the boundary completion process. The original input is two noisy but vertically biased inducing points and a horizontal intervening noise element. The cooperative-competitive exchange triggers transient orthogonal inductions before attenuating all nonvertical elements as it completes the vertical boundary.

7. Textural Segmentation and Grouping. While the importance of the Boundary Contour System is illustrated by its ability to complete individual contours, whether "real" or "illusory," its rules also appear to account for much of the segmentation of textured scenes into grouped regions separated by perceived contours, many of which have no obvious correlates in the optical input. A contour in a pattern of luminances is generally defined as a spatial discontinuity in luminance. While usually sufficient, however, such discontinuities are by no means necessary for sustaining perceived contours. Regions separated by visual contours also occur in the presence of statistical differences in textural qualities such as orientation, shape, density, or color (Beck, 1966a, 1966b, 1972, 1982, 1983; Beck, Prazdny, and Rosenfeld, 1983; Caelli, 1982, 1983; Caelli and Julesz, 1979). Two findings of textural grouping research are especially salient. First, the visual system's segmentation of the scenic input occurs rapidly throughout all regions of that input, in a manner often described as "preattentive." That is, subjects generally describe boundaries in a consistent manner when exposure times are short (under 200 msec) and without prior knowledge of the regions in a display at which boundaries are likely to occur. Thus any theoretical account of boundary extraction for such displays must explain how early "data driven" processes rapidly converge on boundaries wherever they occur.

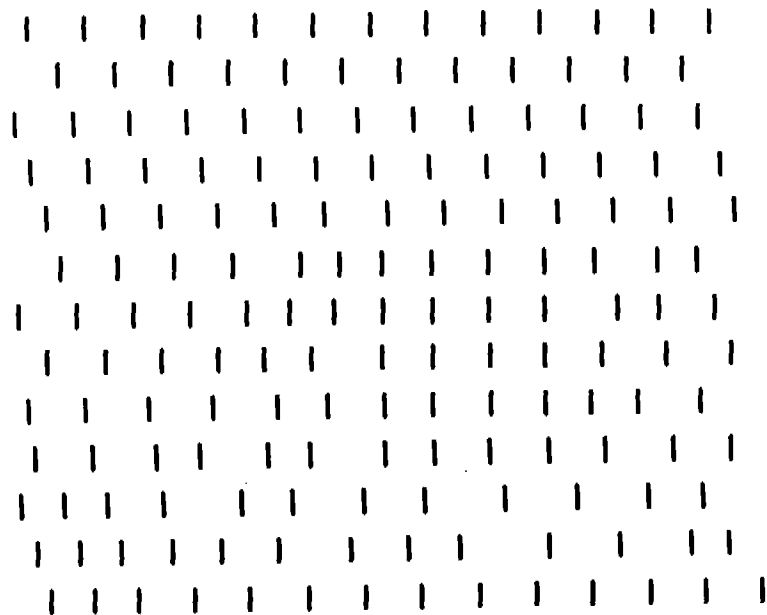
The second finding of the experimental work on textures complicates the implications of the first, however: the textural segmentation process is exquisitely context-sensitive. That is, a given texture element at a given location can be part of a variety of larger groupings, depending on what surrounds it. Indeed, the precise determination even of what acts as an *element* at a given location can depend on patterns at nearby locations. We suggest that a boundary completion process underlies the context-sensitive ability of the visual system to segment and group textured input.

A long line of distinguished research by Jacob Beck and his colleagues has identified variables affecting textural segmentation by the human visual system (Beck, 1983; Beck, Prazdny, and Rosenfeld, 1983). Simple displays like the ones shown in Figure 6 show that the slopes of small elements of color or brightness contrast are a critical determinant of grouping, with regions containing many features with similar slopes tending to group. In particular, if certain of these features are distributed in a regular manner, colinear groupings of these features can become "emergent features," capable of setting one textural region apart from another. A crucial aspect of such emergent features is that the colinear arrangement need not be in line with the directions of the local contrasts. (See Figure 6).

Computer simulations illustrate the ability of the Boundary Contour System to generate perceptual groupings akin to those seen in Figure 6. Numerical parameters were held fixed for all of the simulations; only the input patterns were varied. As the input patterns were moved about, the Boundary Contour System sensed relationships among the inducing elements and generated emergent boundary groupings among them. In all of the simulations, we defined the input patterns to be the output patterns of the oriented receptive fields, as in Figure 4a, since our primary objective was to study the cooperative-competitive feedback exchange. All possible oriented groupings generated inputs to the cooperative-competitive feedback process. Only the favored groupings survived. Thus the ability of the network to suppress the many incorrect local groupings is as important as its ability to choose the correct global grouping.

Figure 7a depicts an array of four vertically oriented input clusters. We call each cluster a *Line* because it represents a caricature of how a field of oriented complex cells respond to a vertical line. Figure 7b displays the equilibrium activities of the cells at the second competitive stage of our model. The length of an oriented line at each position is proportional to the equilibrium activity of a cell whose receptive field is centered at that position with that orientation. The input pattern in Figure 7a possesses a manifest vertical symmetry: Pairs of vertical *Lines* are colinear in the vertical direction, whereas they are spatially out-of-phase in the horizontal direction. The Boundary Contour System senses

(a)



(b)

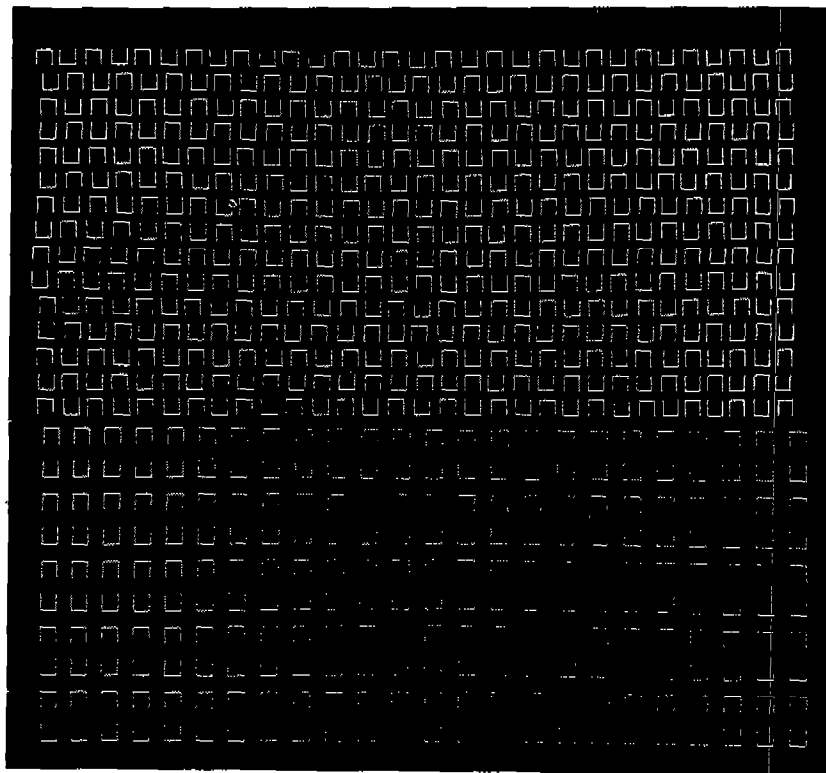


Figure 6: (a) Emergent features. The colinear linking of short line segments into longer segments is an "emergent feature" which sustains textural grouping. Our theory explains how such emergent features can contribute to perceptual grouping even if they are not visible (Gellatly, 1980). (Reprinted from Beck, Prazdny, and Rosenfeld, 1983.) (b) The diagonal grouping at the top of this figure is initiated by differential activation of diagonally oriented receptive fields, despite the absence of any diagonal edges in the image. Horizontal cooperation of signals at the ends of vertical lines generates subjective contours in the bottom half of this figure. (Adapted from Beck, Prazdny, and Rosenfeld, 1983.)

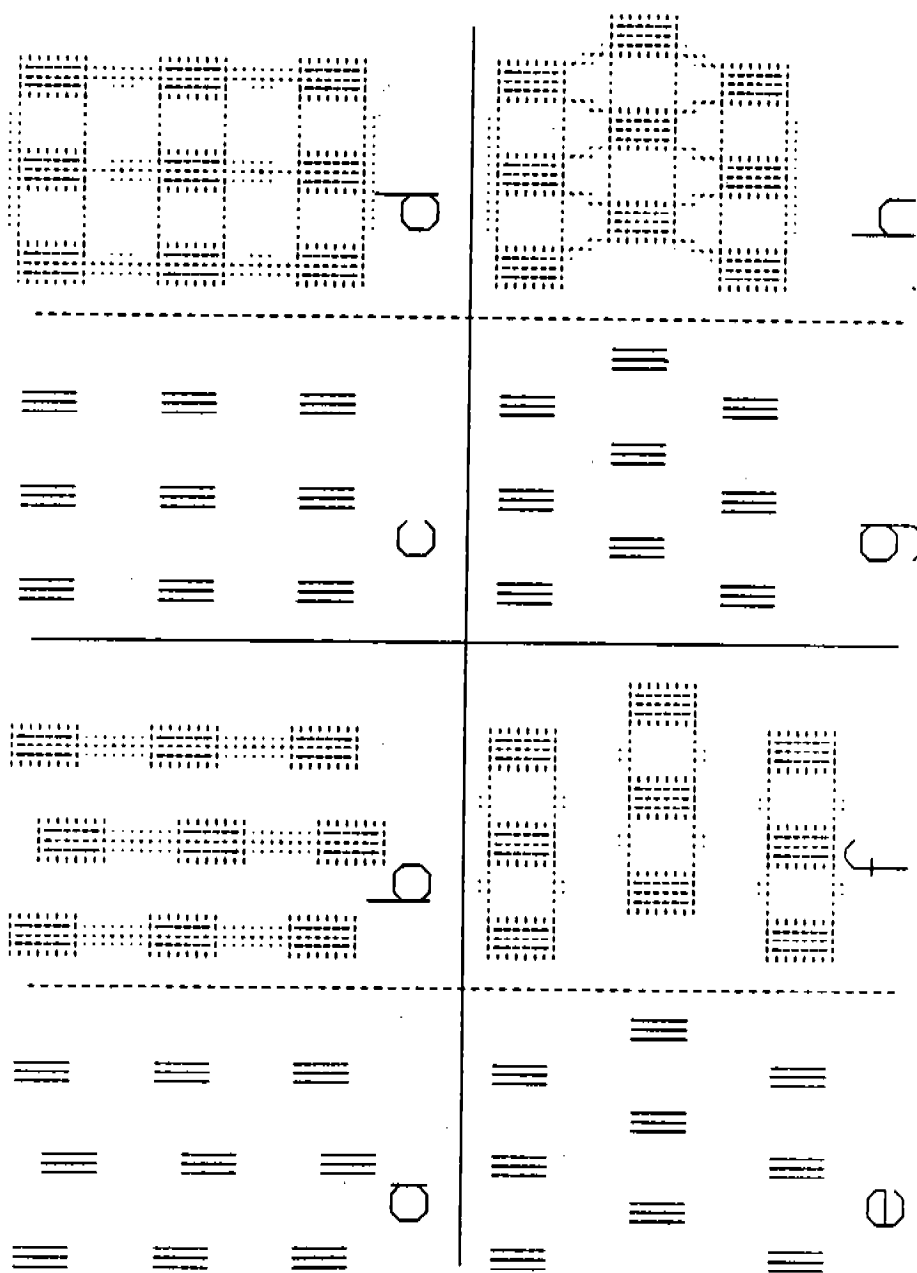


Figure 7: Computer simulations of processes underlying textural grouping: The length of each line segment is proportional to the activation of a network node responsive to one of twelve possible orientations. The dots indicate the positions of inactive cells. Parts (a), (c), (e), and (g) display the activities of oriented cells which input to the cooperative-competitive feedback process. Parts (b), (d), (f), and (h) display equilibrium activities of oriented cells at the competitive stage of the Boundary Contour System. A pairwise comparison of (a) with (b), (c), and (d), and so on indicates the major groupings senses by the network.

this vertical symmetry, and generates emergent vertical lines in Figure 7b. In addition, the Boundary Contour System generates horizontal end cuts at the ends of each Line, which can trap the featural contrasts of each line within the Feature Contour System.

In Figure 7c the input Lines are moved so that pairs of Lines are colinear in the vertical direction and their Line ends are lined up in the horizontal direction. Now both vertical and horizontal groupings are generated in Figure 7d. In Figure 7e the input lines are shifted so that they become non-colinear in a vertical direction, but pairs of their Line ends remain aligned. The vertical symmetry of Figure 7c is hereby broken. Thus in Figure 7f the Boundary Contour System groups the horizontal Line ends, but not the vertical Lines.

Figure 7h depicts a more demanding phenomenon: the emergence of diagonal groupings where no diagonals exist in the input pattern. Figure 7g is generated by bringing the two horizontal rows of vertical Lines closer together until their ends lie within the spatial bandwidth of the cooperative interaction. Figure 7h shows that the Boundary Contour System senses diagonal groupings of the Lines. These diagonal groupings emerge on both microscopic and macroscopic scales. Thus diagonally oriented receptive fields are activated in the emergent boundaries, and these activations, as a whole, group into diagonal bands.

These figures illustrate that the Boundary Contour System behaves like an on-line statistical decision theory, sensing only those groupings of perceptual elements with enough "statistical inertia" to drive its cooperative-competitive feedback exchanges towards a non-zero stable equilibrium configuration. One can interpret the distribution of oriented activities at each input position as being analogous to a local probability distribution, and the final Boundary Contour System pattern as being the global decision that the system reaches and stores based upon all of its local data. In contrast to stochastic relaxation algorithms for boundary detection (Geman and Geman, 1984) wherein a formal temperature parameter is slowly decreased to drive the system towards a minimal energy configuration with boundary enhancing properties, no Boundary Contour System parameter is manipulated by an external agent. The Boundary Contour System internally regulates its own convergence to a coherent configuration via its cooperative-competitive feedback exchanges. The input patterns themselves are the only "external parameters" that are altered in our system. Changing the external input pattern can cause a global switch, or phase transition, in the network, as in Figure 7. In each new mode, the network can maintain a different coherent organization via its cooperative-competitive feedback loops. The network can sustain a large number of different coherent configurations using its nonlinear dissipative dynamics (Prigogine, 1978), rather than a conservative Hamiltonian system.

The present simulations were generated by solving large systems of nonlinear differential equations on a traditional computer. These simulations suggest that many seemingly esoteric Gestalt rules for grouping unambiguous wholes from ambiguous parts can be effected by networks of cells embodying a small number of Boundary Contour System stages. The local intelligence required to implement these network designs is far less than that required of a traditional computer central processing unit. The intelligence of the networks comes from their embodiment of nonclassical geometrical properties that are inherently context-sensitive. An appropriate hardware implementation of the Boundary Contour System would segment and group complex and noisy images in a context-sensitive manner and in real-time.

8. A Synthesis of Nonlinear Dynamics, Parallel Computation, and Image Processing. The deepest conceptual issues raised by the present results concern the choice of perceptual units and design principles. Local computations of scenic elements cannot provide an adequate understanding of visual perception, if only because most luminance changes are discounted as spurious by the human visual system. The Boundary-Feature Trade-Off shows that the visual system is designed in a way that is quite different from any

possible local computational theory. A fertile source of new ideas about parallel computation, image processing, geometry, and statistical mechanics, no less than about perceptual and neural theory, can thus be found in the collective properties of these very large systems of nonlinear differential equations.

REFERENCES

- Beck, J., Perceptual grouping produced by changes in orientation and shape. *Science*, 1966, **154**, 538-540 (a).
- Beck, J., Effect of orientation and of shape similarity on perceptual grouping. *Percept. Psychophys.*, 1966, **1**, 300-302 (b).
- Beck, J., Similarity grouping and peripheral discriminability under uncertainty. *Amer. J. Psych.*, 1972, **85**, 1-19.
- Beck, J., Textural segmentation. In J. Beck (Ed.), **Organization and Representation in Perception**. Hillsdale, NJ: Erlbaum, 1982.
- Beck, J., Textural segmentation, second-order statistics, and textural elements. *Biol. Cybern.*, 1983, **48**, 125-130.
- Beck, J., Prazdny, K., and Rosenfeld, A., A theory of textural segmentation. In J. Beck, B. Hope, and A. Rosenfeld (Eds.), **Human and Machine Vision**. New York: Academic Press, 1983.
- Caelli, T., On discriminating visual textures and images. *Percept. Psychophys.*, 1982, **31**, 149-159.
- Caelli, T., Energy processing and coding factors in texture discrimination and image processing. *Percept. Psychophys.*, 1983, **34**, 349-355.
- Caelli, T. and Julesz, B., Psychophysical evidence for global feature processing in visual texture discrimination. *J. Opt. Soc. Amer.*, 1979, **69**, 675-677.
- Carpenter, G.A. and Grossberg, S., A neural theory of circadian rhythms: The gated pacemaker. *Biol. Cybern.*, 1983, **48**, 35-59.
- Carpenter, G.A. and Grossberg, S., A neural theory of circadian rhythms: Aschoff's rule in diurnal and nocturnal mammals. *Amer. J. Physiol.*, 1984, **247**, R1067-R1082.
- Carpenter, G.A. and Grossberg, S., A neural theory of circadian rhythms: Split rhythms, after-effects, and motivational interactions. *J. Theor. Biol.*, 1985, **113**, 163-223.
- Cohen, M.A. and Grossberg, S., Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Trans.*, 1983, **SMC-13**, 815-826.
- Cohen, M.A. and Grossberg, S., Some global properties of binocular resonances: Disparity matching, filling-in, and figure-ground synthesis. In P. Dodwell and T. Caelli (Eds.), **Figural Synthesis**. Hillsdale, NJ: Erlbaum, 1984 (a).
- Cohen, M.A. and Grossberg, S., Neural dynamics of brightness perception: Features, boundaries, and resonance. *Percept. Psychophys.*, 1984, **36**, 428-456 (b).
- Gellatly, A.R.H., Perception of an illusory triangle with masked inducing figure. *Perception*, 1980, **9**, 599-602.
- Geman, S. and Geman, D., Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 1984, **6**, 721-741.
- Grossberg, S., Competition, decision, and consensus. *J. Math. Anal. Appl.*, 1978, **66**, 470-493.
- Grossberg, S., Biological competition: Decision rules, pattern formation, and oscillations. *Proc. Natl. Acad. Sci.*, 1980, **77**, 2338-2342.

- Grossberg, S., The quantized geometry of visual space: The coherent computation of depth, form, and lightness. *Behav. Brain Sci.*, 1983, **6**, 625-657 (a).
- Grossberg, S., Neural substrates of binocular form perception: Filtering, matching, diffusion, and resonance. In E. Basar, H. Flohr, H. Haken, and A.J. Mandell (Eds.), **Synergetics of the Brain**. New York: Springer-Verlag, 1983 (b).
- Grossberg, S., Outline of a theory of brightness, color, and form perception. In E. Degreef and J. van Buggenhout (Eds.), **Trends in Mathematical Psychology**. Amsterdam: North-Holland, 1984.
- Grossberg, S. and Mingolla, E., Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psych. Rev.*, 1985, **92**, 173-211 (a).
- Grossberg, S. and Mingolla, E., Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Percept. Psychophys.*, 1985, in press (b).
- Helmholtz, H. von, **Physiological Optics**, J.P.C. Southall (Ed.). New York: Dover, 1962.
- Hubel, D.H. and Wiesel, T.N., Functional architecture of macaque monkey visual cortex. *Proc. Roy. Soc. London (B)*, 1977, **198**, 1-59.
- Kanizsa, G., Contours without gradients or cognitive contours? *Ital. J. Psych.*, 1974, **1**, 93-113.
- Kennedy, J.M., Subjective contours, contrast, and assimilation. In C.F. Nodine and D.F. Fisher (Eds.), **Perception and Pictorial Representation**. New York: Praeger, 1979.
- Krauskopf, J., Effect of retinal image stabilization on the appearance of heterochromatic targets. *J. Opt. Soc. Amer.*, 1963, **53**, 741-744.
- Land, E.H., The retinex theory of color vision. *Sci. Amer.*, 1977, **237**, 108-128.
- Prigogine, I., Time, structure, and fluctuations. *Science*, 1978, **201**, 777-786.
- Redies, C. and Spillmann, L., The neon color effect in the Ehrenstein illusion. *Perception*, 1981, **10**, 667-681.
- van Tuijl, H.F.J.M., A new visual illusion: Neonlike color spreading and complementary color induction between subjective contours. *Acta Psych.*, 1975, **39**, 441-445.
- Yarbus, A.L., **Eye Movements and Vision**. New York: Plenum, 1967.

Genie: An Inference Engine with Vulnerability Application

6 May 1985

**BRL VLD AI Research Team
USA Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
301-278-6316**

**Fred Brundick
<fsbrn@ brl.arpa>**

**John Dumer
<dumer@ brl.arpa>**

**Ralph Shear
<shear@ brl.arpa>**

**Timothy Hanratty
<hanratty@ brl.arpa>**

**Paul Tanenbaum
<pjt@ brl.arpa>**

I. INTRODUCTION

A. Background.

Determining the vulnerability of combat vehicles and other systems to the wide range of threats presented in modern warfare can be a daunting task. Although physics, engineering, operations research, and other disciplines offer an invaluable framework for addressing the problem, these more-or-less hard sciences cannot by themselves provide complete solutions. On the contrary, practitioners of vulnerability analysis will assure you that their profession is at least as much an admixture of art, intuition, and educated guesswork as a science.

Traditional computer techniques, although useful, have been unable to master the entire problem. It was to address this problem that the artificial intelligence (AI) community developed the technology that has come to be called *expert systems*. A true expert system is a program that mimics the performance of a human expert in some intellectual endeavor. The archetypal expert system attains this high level of proficiency by embodying the heuristic, informally framed knowledge of the human expert, along with the expert's not-always rigorous methods of reasoning in the subject domain.

We are building expert systems to deal with vulnerability analysis. Toward this end, we have developed a general-purpose mechanism (called an *inference engine* in AI terminology) with which to emulate the behavior of human experts. The project, developed in Franz LISP on a VAX-11/750, is called *Genie*, and this paper offers a presentation of its design, its implementation, and its usage.

As our first application of *Genie*, we are working with Walter Thompson and Steven Polyak of the Aerial Targets Branch, VLD, on an expert system to assist human experts in assessing the vulnerability of turbine jet engines. Many of the examples cited in this paper can be thought of as being taken from such a system, although authenticity of domain details will sometimes suffer for the sake of illustrative clarity.

B. Typical Architecture of an Expert System.

The highest level of organization of an expert system is not complex, as figure 1 illustrates. A basic tenet of the expert system methodology requires separation of domain expertise (whether it be in infectious diseases or turbine engine vulnerability) from the strategies for manipulating and applying that expertise. The resultant modularity greatly facilitates debugging and modifying either subsystem.

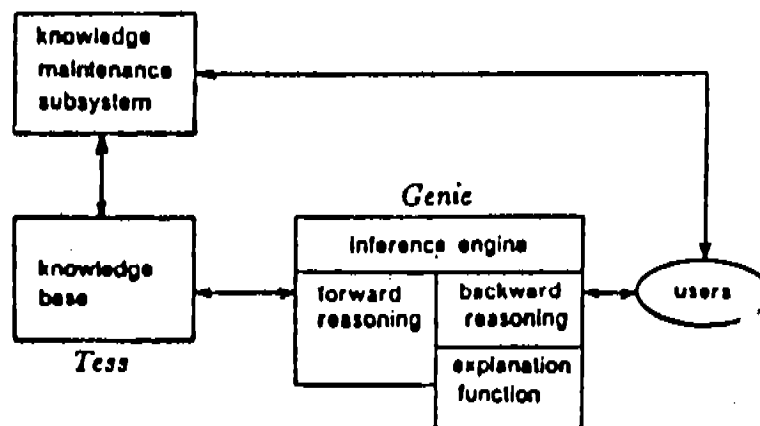


FIGURE 1— System Overview

The first of the subsystems is called the *knowledge base*. It is the mass of expertise that has been collected from one or several human experts. The knowledge base will typically contain both declarative knowledge (e.g. that the specific gravity of steel is 7.8) and procedural knowledge (e.g. that a means of determining the specific gravity of an arbitrary solid is to

calculate the ratio of the solid's density to the density of water). Procedural knowledge is often expressed in rules of propositional implication (e.g. that if a solid is known to be roughly as dense as steel then it can be deduced that the solid's specific gravity is approximately equal to that of steel). The process of constructing a knowledge base, of determining how an expert solves problems and implementing that knowledge on a computer, is called *knowledge engineering*.

The second subsystem is the *inference engine*. In a sense, the inference engine is the program that acts upon the knowledge base as data. Optimally, it would be so well separated from the knowledge base and so general that one could plug in a different knowledge base and thereby create a new expert system in an unrelated domain.

Though it might not be as fundamental, an interface between the system and its users is no less important than the knowledge base or the inference engine. The user interface should allow a non-programmer to modify the knowledge base, obtain consultative assistance, and interrogate the system about the reasons for its behavior, all in language that is clear and natural to the user.

II. REPRESENTATION OF KNOWLEDGE

A key subproblem in modeling the behavior of a human expert is to choose appropriate formats for expressing and storing the expert's knowledge so that a computer can process the knowledge effectively and efficiently. In this section we discuss the approach to knowledge representation that has been employed in building Genie. The discussion is presented in two parts: first we consider the syntax of our representation scheme, the formats available for use; and then we consider their semantic content, what mental constructs these structures represent.

A. Tools

1. Frames. The data structure that is most widely used by Genie is called a *frame*. Frames are a method of organizing information about an object and its properties, together with pointers to other frames describing related objects. A frame is composed of slots, each of which contains a piece of information. Figure 2 illustrates the kinds of information that might be stored in a frame.

Comp.T58	
A Kind Of	compressor
Type	axial
Manufacturer	General Electric
Found In	T58 turbine engine
Pressure Ratio	8.3 : 1
Air Mass Flow	5.6 kg/s
Blade Material	steel
Exit Pressure	calculate: intake pressure times pressure ratio

FIGURE 2.—*Typical Frame*

Of course, if frames are to be used, then special functions will be required to manipulate them. We have implemented a set of routines to retrieve and store information, and to perform other such tasks in the manner described by Winston and Horn¹ and by Roberts and Goldstein².

2. Value Ranges. Arithmetic manipulation of data is a central theme of computer science. Expert systems do not manipulate numbers as often as they do more general symbols, but a means to perform numeric computation is unquestionably among Genie's requirements. Unfortunately, the kinds of numeric knowledge which Genie must handle carry the same imprecision and uncertainty that first motivated the development of expert systems. There are in the literature many approaches to dealing with the problem of imprecise and uncertain knowledge, some of which will be discussed in section VIII below. Another representation problem can turn up even with knowledge that is perfectly precise, crisp, and certain. Some knowledge is inescapably cast as information about sets, vectors, or intervals, rather than scalars.

We have developed a representation format to deal with both of these problems. Using what we call *value ranges* and the routines for manipulating them, one can represent either intervals on the real line or imprecisely known scalars. A value range record consists of six cells which can be labeled as in figure 3.

>	≥	=	≠	≤	<
---	---	---	---	---	---

FIGURE 3.—*Format of a Value Range Record*

Either of the first two cells can be used for a lower bound on the value. Similarly, the last two cells can bound the value above. The \neq cell can contain a set of prohibited values. The $=$ cell can only hold one value; if it is non-empty, then all the other cells must be empty. As an example, the value range in the proposition that

$$0 < x \leq 10, \quad x \text{ not } \in \left\{ \frac{\pi}{2}, 6, 7 \right\} \quad (1)$$

would be represented by the record

0	,		$\left\{ \frac{\pi}{2}, 6, 7 \right\}$	10	
---	---	--	--	----	--

FIGURE 4.—*Inequality (1) Represented as a Value Range*

One way that Genie uses value ranges is in representing rules in the knowledge base. For example, a knowledge base might contain the following rule,

"If thrust of engine < 6000 lbs, or
thrust of engine > 12,000 lbs,
then engine is not J57."

Another use for value ranges is Genie's facility for interpreting a user's answers to its

¹ P. H. Winston and B. K. P. Horn. *LISP*. Addison-Wesley, 1981. pp. 291-301.

² R. B. Roberts and I. P. Goldstein. *The FRL Primer*. Memo No. 409, Artificial Intelligence Laboratory, MIT, 1977.

questions. In answer to the question, "What is the thrust of engine?" the user may enter, "5000 <= thrust < 7500". There are several functions that operate on value ranges, including routines to build a record, to determine if a hypothesized value conflicts with existing knowledge, and to improve the precision of the system's estimate of a value.

B. Semantics

For the most part, Genie manipulates knowledge that is expressed in propositional implications known as *production rules*, or *productions*. A production links one group of propositions, called its *antecedents*, to a second group of propositions, called *conclusions*. The individual propositions are represented by fact frames, and the productions by rule frames. A third type of frame used by Genie is the concept frame. All three frame types have two states. The static state consists of intrinsic information and relations; it is run-time invariant. The dynamic state adds the results of all the manipulations performed during the current execution.

1. **Rules.** In practice, a rule frame often contains just a list of preconditions for applying the rule together with a list of deductions that result from its application. These are stored in the *Ifs* and *Thens* slots, respectively. Since a proposition and its negation describe the same knowledge, we use one fact frame to represent both. Therefore, propositions appearing in rules must be marked to specify which sense should be used. As an example, consider rule6, whose frame appears in figure 5a. It states that if fact19 and fact20 are known to be true and fact3 is known to be false then fact21 is true and can be added to the knowledge base.

Once the three preconditions are met and the deduction is made, rule6 takes the form given in figure 5b. In this case, rule6 is said to have *fired*. If, however, fact20 were determined to be false, the rule would not fire and nothing would be learned about fact21's truth value. This situation is illustrated in figure 5c.

Flexibility in specifying a rule's preconditions is provided by the *must-have* mechanism shown in figure 5d. The rule88 frame contains three antecedents. The *Must-have* slot indicates that if any two of the antecedents can be determined to hold, then the rule can fire. The default, for rules without *Must-have* slots, requires all the preconditions to be met.

2. **Facts.** The fact is the basic semantic building block in a Genie knowledge base. Rules are built up from them, and the inference engine attempts to deduce or verify them. When the user is asked a question, it is in order to acquire new facts. And the system's ultimate answer is some fact. The internal representation of facts is illustrated in figure 6.

The frame fact17 (figure 6a) represents a proposition about a compressor. The English-language statement of that proposition is found in the *Stmnt* slot. The *Default* slot is available to specify which sense of a proposition (viz. its affirmation or its negation) should be assumed in the event that its truth value cannot be determined directly. The *Ifs-of* slot lists all the rules in the knowledge base whose application depends upon the truth value of fact17. In particular, fact17 is in the *Ifs* slots of rules five, twenty-seven, and twenty-eight. The *Thens-of* slot, on the other hand, lists all the rules that, if applied, will determine fact17's truth value: fact17 is in the *Thens* slots of rules 16 and 105. This linking of facts to the rules that use them speeds execution and facilitates explaining system behavior to the user.

The *Xor* slot of fact17 contains the name of a group of mutually exclusive propositions. If, by whatever means, fact17 is determined to be true, then Genie will conclude that all the other facts in xor3 are false. This feature is a handy way to model hierarchies of disjoint classes of objects (e.g. Whether a target is of RHA, mild steel, or aluminum). It also provides a means of representing the relation between antonyms, as for example, whether compressor vane geometry is variable or fixed.

Suppose Genie was running and applied rule16. As a result, it would amass dynamic knowledge about the truth value of various facts. If rule16's assertion about fact17 were in the affirmative, then fact17 would be modified to the form shown in figure 6b. The *How* slot could be checked at a later time to determine the context in which fact17's truth was determined.

rule6	
Ifs	(fact19 1) (fact20 1) (fact3 0)
Thens	(fact21 1)

(a)

rule6	
Ifs	(fact19 1) (fact20 1) (fact3 0)
Thens	(fact21 1)
Status	Fired

(b)

rule6	
Ifs	(fact19 1) (fact20 1) (fact3 0)
Thens	(fact21 1)
Status	Failed
Culprit	fact20

(c)

rule88	
Ifs	(fact15 1) (fact32 0) (fact33 1)
Thens	(fact34 1) (fact5 0)
Must-have	2

(d)

FIGURE 5.—Sample Rule Frames.

- (a) The static version of a simple rule.
 (b) The dynamic version of a successful application of (a).
 (c) A dynamic version of a failed application of (a).
 (d) A must-have rule.

Propositions concerning numeric knowledge can be represented using what we call *arithmetic facts*, as illustrated in figure 6c. The **Arithmetic** slot of an arithmetic fact frame has three sub-cells (called *facets* in frame terminology), which provide the links by which to confirm or disprove fact23. According to the **concept** and **attr** facets, fact23 concerns the attribute called speed of an object called spool. Genie determined that fact23 was false by looking in the **spool** frame to compare what it knew about spool speed with the relation stored in the **relat** facet of fact23.

3. **Concepts.** Rule and fact frames provide a significant increase in deductive speed, but they are little more than an extension of the production-rule approach to building expert systems. One of the major advantages of production rules over other semantic structures is their modularity. Rules represent small pieces of knowledge and can be added to a knowledge base or modified easily, and undesired side-effects are less common than with more complex

fact17	
Stmt	Stator vanes are aluminum
Ifs~of	rule5 rule 27 rule 28
Thens~of	rule16 rule105
Xor	xor3
Default	Stator vanes are not aluminum

(a)

fact17	
Stmt	Stator vanes are aluminum
Ifs~of	rule5 rule 27 rule 28
Thens~of	rule16 rule105
Xor	xor3
Default	Stator vanes are not aluminum
Truth	True
How	Deduced using rule16

(b)

fact23		
Stmt	Speed of spool GT 12000	
Ifs~of	rule18 rule120	
Arithmetic	concept	spool
	attr	speed
	relat	GT 12000
Truth	False	
How	Num-Relat	

(c)

FIGURE 6.—Sample Fact Frames.

- (a) The static version of a simple fact. (b) A dynamic version of (a).
(c) A dynamic version of an arithmetic fact.

structures. Unfortunately, though, the modularity of production rules also represents one of their most serious drawbacks. Because a knowledge base made up of rules has such a fine granularity, it is difficult to ascertain high-level patterns and order in the knowledge. This problem is especially serious when a user tries to understand the system's lines of reasoning, or when a student tries to acquire the expertise inherent in the knowledge base.³

As a further step towards representing the order in an expert's knowledge, Genie uses concept frames to group information. In the discussion of arithmetic facts, above, we saw one

³ A. Barr and E. A. Feigenbaum, eds. *The Handbook of Artificial Intelligence*, vol. 1. William Kaufmann, Inc., 1981. pp. 193-194.

circumstance in which concept frames are used. Proceeding with that example, we shall trace Genie's process of determining the truth value of fact23. The system's knowledge about several of the spool's attributes are stored in **spool**, which might take the form shown in figure 7.

spool							
speed	value~range			9590			
	used~by	fact23	fact50	fact82			
	asked	Yes					
	value	9590					
length	value~range						
	used~by	fact40	fact95				
mass	value~range	0				500	
	used~by	fact43	fact44	fact60	fact99		
	asked	Yes					

FIGURE 7.—Sample Concept Frame.

Originally, there was insufficient information in the **speed** slot of **spool** to determine whether fact23 held. So Genie asked the user, "What is the speed of spool?" The user's response was presumably the scalar value 9590, since that is what the **value~range** facet indicates. When the =cell of a **value~range** is non-empty, its contents is also stored in the **value** facet.

Neither fact40 nor fact95 has been needed yet, since the **length** slot retains its static configuration. But one of the four propositions about spool mass was needed, since the **asked** facet of **mass** is full. If one of the three remaining facts asserted that spool mass equaled 200, then Genie would be incapable of determining that fact's truth value by direct calculation. This is because Genie will only request input of a given parameter once, on the assumption that the user will have given his best estimate immediately.

III. REASONING METHODS

The basic intent in developing expert systems is to enable a computer to reason as a human expert does. By this we mean *to achieve expert performance*, since our goal is not so much modeling the expert's behavior as modeling the results of that behavior. A parallel can be drawn with the expert system's representation of knowledge: one seeks a format that is isomorphic to the one used by the expert's brain, but replication is neither necessary nor possible. For example, we do not assert that frames or value-range records exist in the human brain. But neither are silicon chips identical to cortical neurons. In the same way, while formal logic is seldom applied by humans to real-world problems, its spirit can be fundamentally useful in developing expert systems.

Given a body of formal assertions and implications, or propositions and production rules, there are basically two possible strategies for using them. One strategy focuses on the rules' antecedents, the other on their conclusions. They are called *forward* and *backward chaining*, respectively, and will be explained below.

A. Forward Chaining

In forward chaining, one compares the facts in the knowledge base with the antecedents of the various rules, trying to fire any rules possible. When a rule fires, its conclusions are added to the knowledge base and can potentially trigger other rules. Because attention is

focused on matching antecedents against the facts that are known, this method is also called *data-driven reasoning*.

Forward chaining is often extremely useful in real-time applications. In these settings, the rules often represent event/response or condition/action knowledge. Robotics and industrial process control are examples.

B. Backward Chaining

The second strategy is somewhat more complicated. In backward chaining one starts by considering a goal, in this case a fact whose truth value is desired. The key step is to find a way of determining the truth value, which usually means finding a rule that draws a conclusion about the fact. If such a rule exists, then one can reformulate the original problem as the determination of the truth values of each of the rule's antecedents. This reformulation of problems into subproblems is iterated until each of the subproblems can be solved directly, either by finding facts in the knowledge base, or by asking questions of the user. For this reason, backward chaining is also called *goal-driven reasoning*.

C. Genie's Approach

Reasoning in the current version of Genie has a dual nature. From a holistic viewpoint, the control strategy is entirely goal-driven. Equally valid, however, is the reductionist point of view that all the system's deductions are made in data-driven mode. The two arguments are presented here.

A Genie knowledge base must contain at least one fact tagged as a *hypothesis*, or top-level goal. The inference engine records all the hypotheses and tries to verify each one in succession. Given a hypothesis, Genie looks at the **Thens~of** slot in its fact frame. This provides a list of rules that could potentially confirm the hypothesis. These rules' **Ifs** slots specify other facts which Genie treats as subgoals. The subgoals will generally have additional rules in their own **Thens~of** slots, and so on. Following all these **Thens~of** and **Ifs** links as far as they lead would generate a tree structure with facts as nodes and rules as edges. The leaves of this tree — facts that cannot be deduced from any rule in the knowledge base — constitute the information Genie must request from the user, and are consequently the simplest possible subgoals.

Consider the lowest level of deduction in this process: a rule whose antecedents are all leaves. A question will be asked for each fact, the user's answers being added to the knowledge base in a process called *memorization*. Assuming that the conditions specified by the rule's antecedents conform to the answers, these subgoals are all achieved. So the rule fires, causing its conclusions to be memorized, and a larger fraction of the problem has been solved. The procedure continues in this fashion.

Not all the rules will fire, of course. A rule fails if its antecedents do not conform to the circumstances of the present run. When this happens, the desired conclusion must be achieved through other means. If the fact has untried rules in its **Thens~of**, they will be tried. If not, the fact frame will be searched for a **Default**. If all the rules that might confirm a hypothesis fail, then the hypothesis is discarded, and another one tried.

Whereas backward chaining imposes order on Genie's performance, it is forward chaining that actually deduces facts. Every time a fact is memorized, whether it was given by the user or deduced from some rule, forward chaining is performed on it. The fact's **Ifs~of** slot lists the rules that require it. Each one of these rules that has not already fired or failed is considered. If any rule's antecedent requires the wrong truth value for the fact, then the rule fails. However, if the newly memorized fact conforms with the only unfulfilled precondition of a rule, then that rule will fire. Genie will then forward chain on each of the rule's conclusions in turn.

The double linking of facts and rules makes Genie's knowledge representation scheme very flexible. Backward chaining is achieved by following the **Thens~of** and **Ifs** links. Similarly, following **Ifs~of** and **Thens** links accomplishes forward chaining. Thus, one representation of a rule can be used for both strategies. This capability is a very important one, since

using only one or the other strategy has serious drawbacks. A system that only forward chains often seems to behave erratically, jumping around the knowledge base. Furthermore, a purely data-driven system cannot even ask the user any questions, since it can only passively obtain facts and apply rules to them. On the other hand, a strictly backward chaining system draws only those conclusions that are of immediate use. So it will miss drawing conclusions that are supported by its knowledge but do not lie in the direct path of its current task.

In summary, backward chaining causes Genie's reasoning to be directed from the broadest, most general conclusions, through ever more specific facts. This top-down behavior creates the impression that the system is acting purposefully. Within that context, forward chaining ensures that deductions are made as soon as the necessary knowledge is acquired.

IV. USER INTERFACE

The two previous sections discussed Genie's layout and performance from an internal viewpoint. Here we shall describe the face that Genie shows to humans who interact with it at a computer terminal. First we consider the outermost layer of the program that mediates communication between the user and Genie. Then the specific modes of giving input to the system are addressed. Finally, we discuss Genie's ability to use in its communication something approaching normal English grammar.

A. Driver

Upon invoking Genie, one interacts with a function, called *Driver*, that directs Genie's operation. Anyone who invokes Driver will be categorized into one of three access classes: user, rule-writer, or programmer. The user class is the broadest. It includes those who use the system for production runs. Driver protects general users and Genie from one another, by restricting both the commands the user may use and the actions Genie may perform for the user. The rule-writer class is intended to include the knowledge engineers who maintain the system. A rule writer may execute any of the commands available to the general user. In addition, the rule writer may modify the knowledge base and dig more deeply into Genie's internal mechanisms to determine what the system is doing and why. The programmer class is the least restrictive. Programmers may execute any of the commands available to the rule writer plus several lower-level commands to debug the inference engine and peripheral components.

Driver understands a number of commands, the most important of which is *run*, the request to start up the expert system. One can also instruct Driver to show all the rules that have been applied, or all the facts or numerical values that have been derived so far in the current run. It is also possible to view rules and facts, either in an English form or in a frame form that is more like their internal representations. Concept frames can be displayed in similar fashion.

B. Run-Time Input Functions

When Genie requires information from the user, there are three schemes by which it can request it. The first and simplest of these is the simple yes/no question. The user may respond to a yes/no question by typing one of a number of responses, each of which is equivalent to one of the responses "yes", "no", and "unknown". The three canonical responses have the effect of declaring the fact (as asked) to be true, false, or indeterminate, respectively.

The second scheme for requesting input is the menu. A menu displays a question and several numbered responses. Each of the responses represents a fact, and all the facts in a given menu form a mutually exclusive set (i.e. they are in an xor list). The user simply types in the number of the desired response. Genie then concludes that the chosen fact is true and that all others in the menu are false.

The third mode of input is the numeric question. Whenever Genie requires an arithmetic fact, rather than simply request verification of that fact, it asks for the numeric value itself. The user may respond to a numeric question in a fairly flexible algebraic notation. Possible responses include "5", " $z > 3.14$ ", and " $0 \leq z < 100$ ". Once some value-range has been stored for the numeric value, all the facts that require the value are checked to determine their truths. It is assumed that the user's input was the best estimate he had of the requested value.

In addition to the permissible answers, all three question modes also allow immediate interrogation about the context in which the question is being asked. Depending on one's access class, one may type "why", to determine why the requested information is needed, "rule", to be shown the rule that is currently being tested, and "hyp", to be shown the hypothesis currently under consideration.

C. Grammar

Genie performs all its reasoning on coded representations of the domain information in the knowledge base. Generally, when it speaks to the user it must translate rules and other structures into English. To do this Genie depends on a simple yet effective pseudo-English grammar. This grammar is used in compiling the knowledge base, and then again whenever pieces of knowledge are displayed to the user.

The primary piece of knowledge that must be formulated into English is the fact. Since a proposition's assertion and its negation represent the same knowledge — in the sense that the truth value of either one follows immediately from that of the other — a single fact frame is used to represent both. Each fact frame contains a Stmt slot, the foundation of the formulation process, in which is stored an affirmative English-language statement of the proposition.

So, both for building fact frames and for displaying knowledge and asking questions, the grammatical requirements are: the means to determine the sense of a statement (affirmative or negative), to switch the sense of a statement, and to turn the statement into a question. Genie accomplishes these tasks through a simple scheme that matches statements against grammatical patterns. Given the fact statement, "Compressor has driver rings", Genie will create the question, "Does compressor have driver rings?". It can complement the fact statement, "Engine is fully encased", yielding "Engine is not fully encased."

V. DATA DIGESTION

It has been a fundamental design goal that Genie should be able to perform all of its transactions with humans in nearly natural language. When a new rule is added to a Genie knowledge base, the first form that it takes is an English-language statement. But because Genie cannot actually reason using natural language directly, there is little justification for using English strings as the medium of knowledge representation within the system. The natural-language components of Genie must be able to switch between the internal representation and statements in English.

Strong arguments *against* storing facts as strings of English words can be made from considerations of efficiency. First, it is redundant and wasteful of space to store a lengthy statement of a fact in every rule that contains the fact. But more importantly, representing the abstract object known as a fact, with all its semantic and contextual baggage, as a mere string of words is unacceptably limiting. Both goal- and data-driven reasoning require the pairing up of antecedents and conclusions of various rules. To do this with the scanty rule representation that provides only word strings requires sequentially checking every rule in the knowledge base, performing a time-consuming word-for-word check to see if facts match. Negated facts cause additional headaches.

For these reasons, Genie includes a module that preprocesses its knowledge base. Internally, rules and facts are represented by frames and referred to by unique names called *code*

symbols (like "rule24" and "fact7"). The key function in the data digestion process is called *build-frames*. It reads the knowledge-base input file and puts all the static knowledge into Genie's internal representation formats. As an example, the rule frame shown in figure 5a might have been created by build-frames from the input in figure 8.

```
(IF (there are struts at the front of the engine)
    (there are two flanges near the front of the engine)
    (flanges are not extremely close together)
  THEN
    (engine has a front frame))
```

FIGURE 8.—Sample Input Rule.

The data digestion process creates rule, fact, and concept frames where necessary, recording them in tables so that, for instance, occurrences of a given fact in subsequent rules will be referred to by the same fact code. So the only searching that Genie must do is in translating English statements into code symbols, and this is done all at once during data digestion. Other steps in data digestion are adding the links between rule and fact frames to allow chaining, and linking all the facts in each xor list. Appendix C provides a specification of the structures permissible in the knowledge-base input file.

When an end user encounters Genie, the knowledge base has already been preprocessed. The improvement in performance obtained through data digestion is marked. In a pure production system the mean time required to make one inference grows linearly as the size of the knowledge base is increased. This is so because backward chaining cannot be performed without searching the entire knowledge base for relevant rules. Digesting the rules as is done in Genie speeds the process considerably, and the time per inference is independent of the size of the knowledge base.⁴

VI. EXPLANATION FACILITIES

A common observation among those who have designed expert systems is that a system must be accessible or it will not be used, because few will feel confident accepting the output of a black box. The program should be able to provide information about its reasoning and justify particular inferences when the user requests it to do so.⁵ Genie has several facilities to provide this kind of explanation of its behavior to the user.

The simplest mechanism available to the user is the straight-forward dump of current knowledge. This can take any of the following forms: listing all the facts whose **Truths** are known, listing all the rules whose **Statuses** are known, and listing all the concept attributes for which some **value-range** is known. A related mechanism is the "find" command, which finds all facts whose **Stmts** contain specified words.

The "show" command can be used to display a rule or a fact — either in English or frame form — or a concept frame. This is often useful in combination with "find".

The highest-level interrogation commands currently available to the user are "how" and "why". An example of the former is "how fact17", meaning, "By what means did you determine the truth value of fact17?". To answer the question, Genie checks the **How** slot in the fact frame, and prints an explanation of its justification for concluding the **Truth** of the fact.

⁴ K. Niwa, K. Sasaki, and H. Ihara. "An Experimental Comparison of Knowledge Representation Schemes". *AI Magazine*, Summer 1984. pp. 29-36.

⁵ B. G. Buchanan and E. H. Shortliffe, eds. *Rule-Based Expert Systems*. Addison-Wesley, 1984. pp. 58-59.

The "why" command asks the question, "To what end did you need that fact?" It can be used to determine Genie's motivation for asking a question. In response, Genie displays the rule that it is currently attempting to fire.

Taken together, these commands allow one to discover the system's lines of reasoning. This is useful for the rule writer in ensuring that rules interact to produce the intended conclusions. It is also useful for the end user in deciding whether to accept the system's conclusions and the system itself.

VII. FUTURE WORK

There are several changes underway to improve both Genie's internal processes and its man/machine interface. Included in these are the system's interfaces with the rule writer and the user, uncertainty, higher-level organization of knowledge, more natural user interface.

A. Uncertainty

There are many kinds of uncertainty inherent in any real-world problem. The omnipresence of uncertainty and inexactitude is made even more bothersome by their intractability. Reasoning effectively in the face of these obstacles is among the most challenging problems in AI.

Among the types of uncertainty often encountered are simple probability of a proposition (e.g. "There is a 75 percent probability that the fragment will perforate the combustor housing."), fuzziness of a proposition (e.g. "The power turbine is very rugged."), rule strength (i.e. the extent to which the rule is applicable), and rule reliability (when knowledge is synthesized from several experts).

We have considered adopting the certainty factor technique used in the MYCIN project. A certainty factor is a scalar that is associated with a proposition and reflects the proposition's probability. Conclusions of rules contain certainty factors, so one rule might be said to provide stronger or weaker evidence than another rule with the same conclusion. However, parametric studies indicate that MYCIN's results are fairly insensitive to the choice of certainty factors, so this method may not be highly useful.

Another enticing tool is the theory of fuzzy sets, introduced by Lotfi Zadeh.^{6,7,8} Fuzzy set theory promises to address many types of uncertainty. We have begun working with the Joint BRL/AMSAA Working Group on Fuzzy Sets to apply this approach to our work.

B. Higher-Level Organization of Knowledge

The exclusive use of rules to represent domain knowledge makes it impossible to capture any but the simplest patterns in the knowledge. More abstract organization can only be represented by more complex structures. In order for Genie to reason more powerfully, explain its behavior at a suitably high level, and be appropriate as a teaching tool, it must have a fair level of abstraction. This consideration is central to the enhancement of Genie's performance.

The first two steps toward higher-level organization were the representation of rules and

⁶ L. A. Zadeh. "Fuzzy Sets," *Information Control*, vol. 8, 1965, pp. 338-353.

⁷ L. A. Zadeh. "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. smc-3, no. 1, January 1973, pp. 28-44.

⁸ L. A. Zadeh. *The Role of Fuzzy Logic in the Management of Uncertainty in Expert Systems*. Memorandum No. UCB/ERL M83/41, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, 1983.

facts as frames, and the use of concept frames. Concept frames are currently used only for numeric values in arithmetic facts. The next step is probably to use a more robust concept frame, like that found in Lenat's program, AM.^{9,10} Thus, in a future incarnation Genie might build concept frames for every "object" mentioned in the rule base, so it could fulfill requests like "Show every rule that mentions annular combustors".

Of course, implementing this capability would require that the input rules be analyzed using a much more powerful grammar than the one with which Genie is currently endowed. Genie cannot now be said to understand its knowledge base in abstract terms: it breaks rules into facts and very effectively handles the relations among them, but on fact statements it only performs surface-level manipulations. Clearly, a deeper understanding would require more intelligence in interpreting the knowledge.

C. Rule Writer

As a rule base grows, keeping it free of conflicts, contradictions, and overlaps becomes extremely difficult. While building individual rules is clear and straight-forward, ensuring that the integration of hundreds of rules produces the desired results can be a problem. It would be a great advantage to have a component that helped the knowledge engineer manage the development and maintenance of the knowledge base, facilitating addition and debugging of rules.

Like most of the enhancements discussed here, a rule writing tool's utility depends on its intelligence. For example, even the ability to recognize when two similar facts have related meanings is difficult to automate. A first-pass rule writer could be made to compile rules into the knowledge base as they were entered. Such a program would be able to determine, just as Genie does now, which rules use given facts.

D. More Natural User Interface

Genie ought to be able to converse with the user in something closer to normal English. This is tied in with increasing the order in the knowledge base, since sophisticated statements are built from and reflect elaborate knowledge structures.

Another major improvement in the user interface will be possible when Genie is moved to a LISP machine in the near future. These single-user work stations provide a remarkably powerful environment for both development and production runs. A combination of mouse, windows, pop-up menus, and high-resolution graphics will make communication simple and quite fast.

⁹ D. Lenat. *AM: An artificial intelligence approach to discovery in mathematics as heuristic search*. SAIL AIM-286, Stanford Artificial Intelligence Laboratory, 1976.

¹⁰ D. A. Waterman and F. Hayes-Roth, eds. *Pattern-Directed Inference Systems*. Academic Press, Inc., 1978. pp. 30-33.

ANALYSIS OF GRADIENT CHANGE THRESHOLDS IN THE
DETECTION OF EDGES OF OBJECTS FROM RANGE DATA

C. N. Shen and R. L. Racicot
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189-5000

ABSTRACT. The detection of the gradient change for an object or an obstacle can be analyzed. The ratio of the magnitude of the gradient change to measurement noise is related to the miss and false alarm of the system. The spacing of the measurements also affects the detection threshold of this gradient change.

I. INTRODUCTION. A laser range finder is installed on the top of a mast attached to a land vehicle or a helicopter. This laser range finder measures the distance between itself and the ground points on a terrain with obstacles. For segmentation, the edges of the boulder and crater must be estimated. In determining the near edges of a boulder or the far edges of a crater, it is necessary to locate the first differences of the slopes of the terrain. This is equivalent to finding the second differences of the terrain range points. The threshold values of the second differences can be in terms of an angle in a vertical plane. The probabilities of miss and false alarm can be ascertained depending on the method of estimation and the scanning scheme.

II. THE LAPLACIAN METHOD. The Laplacian Method considers cross-sections of terrain and looks for changes in slope in the azimuth or radial direction. Since we are only considering one-dimensional problems, we can write for the measurement equation

$$z_i = d_i + v_i \quad (1)$$

The change in slope is estimated by computing the following sufficient statistic:

$$\Delta s_i = z_{i+1} - 2z_i + z_{i-1} \quad (2)$$

This is a digital approximation of the second derivative, or second difference, of the range.

Due to the presence of measurement noise in the calculation of s_i , it is impossible to discern with absolute certainty whether a change of slope exists. The Neyman-Pearson criterion provides a decision rule which we may use to accurately detect edges of obstacles with a known probability of making an error. To produce the decision rule, first we must compute the variance of s_i as follows. We assume that the noise components v_i of the measurements are independent Gaussian random variables with zero mean and a variance of σ^2 . Then, since z_{i+1} , z_i , and z_{i-1} are independent, Eqs. (1) and (2) yield

$$\begin{aligned}
\text{var}(s_i) &= \text{var}(z_{i+1} - 2z_i + z_{i+1}) \\
&= \text{var}(z_{i+1}) + \text{var}(-2z_i) + \text{var}(z_{i+1}) \\
&= \sigma^2 + 4\sigma^2 + \sigma^2 = 6\sigma^2
\end{aligned} \tag{3}$$

Because s_i is a scalar random variable, the Neyman-Pearson criterion provides a decision rule identical to that derived using hypothesis testing. The desired decision rule is

$$\begin{aligned}
\text{DECISION}_i &= \begin{array}{ll} \text{no signal} & \text{if } -T < s_i < T \\ \text{presence of signal} & \text{if } s_i \text{ is otherwise} \end{array}
\end{aligned} \tag{4}$$

where T is the threshold in the decision process which can be determined from the equations

$$P_F = 2\phi[-T/\sqrt{6}\sigma^2] \tag{5}$$

$$P_M = \phi[(T-u^*)/\sqrt{6}\sigma^2] \tag{6}$$

where

$$\phi(z) = 1/\sqrt{2\pi} \int_{-\infty}^z e^{-\alpha^2/2} d\alpha \tag{7}$$

It is noted that the magnitude of s_i is not estimated by the Laplacian Method. The quantity u^* is called the "minimum detectable change of slope," since it is the smallest change of slope that can be detected with a miss probability of P_M or lower. The quantity P_F indicates the probability of a false alarm. The miss probability P_M is the probability of not detecting a true change of slope equal to u^* .

Typically, the standard deviation \sqrt{R} is a known system parameter and u^* is chosen so that suitable values of P_F and P_M can be obtained. Table 1 shows the trade-off of P_F vs. P_M for different values of the ratio u^*/\sqrt{R} . These values are computed using Eqs. (5) and (6).

TABLE 1. FALSE ALARM PROBABILITIES AND MISS PROBABILITIES FOR VARIOUS T/\sqrt{R} AND u^*/\sqrt{R}

T/\sqrt{R}	P_F	P_M For $u^*/\sqrt{R}=2$	P_M For $u^*/\sqrt{R}=3$	P_M For $u^*/\sqrt{R}=4$	P_M For $u^*/\sqrt{R}=5$	P_M For $u^*/\sqrt{R}=6$
1.000	.3174	.1587	.0227	.0013	.0000	.0000
1.500	.1336	.3085	.0668	.0062	.0002	.0000
1.679	.0932	.3741	.0932	.0102	.0005	.0000
2.000	.0456	.5000	.1587	.0227	.0013	.0000
2.146	.0319	.5580	.1966	.0319	.0022	.0001
2.500	.0124	.6915	.3085	.0668	.0062	.0002
3.000	.0026	.8413	.5000	.1587	.0227	.0013
3.500	.0003	.9332	.6915	.3085	.0668	.0062
4.000	.0000	.9772	.8413	.5000	.1587	.0227

III. THE RAPID ESTIMATION SCHEME AND COMPARISON OF ERROR PROBABILITIES. The variance of s_1 for the Laplacian Method is usually too high when a change of slope occurs during scanning. An adaptive method called Rapid Estimation Scheme is used instead to keep the variance lower, thus also reducing the probabilities of false alarm and miss. A discussion of the Second Residual Method for the Rapid Estimation Scheme is given in the Appendix.

The different expressions for the error probabilities of the two methods are given in Eqs. (5) and (6) and reiterated in Table 2. We see that the expressions for both methods are the same except that $\sqrt{6\sigma^2}$ appears with the Laplacian Method, whereas $\sqrt{s_{1+2}}$ appears with the Second Residual Method.

TABLE 2. ERROR PROBABILITIES

Method	Probability of False Alarm	Probability of Miss
Laplacian	$P_L = 2\phi[-T/\sqrt{6\sigma^2}]$ F	$P_L = \phi[(T-u^*)/\sqrt{6\sigma^2}]$ M
Second Residual	$P_S = 2\phi[-T/\sqrt{s_{1+2}}]$ F	$P_S = \phi[(T-u^*)/\sqrt{s_{1+2}}]$ M

We know that the Second Residual variance is less than or equal to the Laplacian variance:

$$\sqrt{s_{1+2}} \leq \sqrt{6\sigma^2} \quad (8)$$

Since $\phi(z)$ is a monotonically increasing function and has a negative argument in all the expressions in Table 2 (we assume $T-u^* < 0$), we can conclude that

$$P_S^F \text{ (Second Residual)} \leq P_L^F \text{ (Laplacian)} \quad (9)$$

$$P_S^M \text{ (Second Residual)} \leq P_L^M \text{ (Laplacian)} \quad (10)$$

Figure 1 illustrates how the smaller Second Residual variance leads to smaller error probabilities.

With smaller error probabilities, we expect the Second Residual Method will have a better performance over the Laplacian Method.

IV. MINIMUM SLOPES OF DETECTABLE OBSTACLES. In this section we will determine the threshold in edging of gradient changes of obstacles. These may be detected using the Second Residual or the Laplacian Method.

Obstacles may or may not be detectable depending on the change of slopes of the terrain at its edges. It will be shown here that the minimum detectable slope change depends upon a given group of parameters.

Figure 2 shows a diagram of an obstacle. C and B are consecutive points where laser beams emanating from laser range finder bounce off the terrain.

The slope of the obstacle is $\tan \theta$, since the slope between C and B is zero. The mast height is b , the height at which the laser range finder is located. Point D is the estimate of the range based on the data at point B and previous points. Thus, the range of point C is measured as z_{i+2} and point D lies at a range of \bar{d}_{i+2} , where $\bar{d}_{i+2} = HF_{i+1} x^*_{i+1}$ is the prediction of the range d_{i+2} from previous data. We see that the distance from C to D is the residue, r_{i+2} . It is apparent that the expected value of r_{i+2} takes its minimum value, u^* . From the geometry, the quantity $\tan \theta$ can be computed by finding x/y . These quantities are found as follows:

$$x = u^* \sin \beta \quad (11)$$

$$y = \Delta \rho - u^* \cos \beta \quad (12)$$

then

$$\tan \theta = x/y = (u^* \sin \beta) / (\Delta \rho - u^* \cos \beta) \quad (13)$$

By extending the above idea, we have two slopes, AB and BC, instead of one in the previous cases. The angle γ is the difference of the angle θ for slope BA and the angle ϕ for slope CB, as shown in Figure 3. Now the slope of γ becomes

$$\tan \gamma = \frac{a}{b}, \quad (\gamma = \theta - \phi) \quad (14)$$

where

$$a = u^* \sin (\beta + \phi) \quad (15)$$

and

$$\begin{aligned} b &= (\Delta \rho) \cos \phi - (\Delta \rho) \sin \phi \cot (\beta + \phi) - u^* \cos (\beta + \phi) \\ &= (\Delta \rho) \frac{\sin (\beta + \phi) \cos \phi - \cos (\beta + \phi) \sin \phi}{\sin (\beta + \phi)} - u^* \cos (\beta + \phi) \\ &= \frac{(\Delta \rho) \sin \beta}{\sin (\beta + \phi)} - u^* \cos (\beta + \phi) \end{aligned}$$

$$= (\Delta \rho) \sin \beta \sin (\beta + \phi) \left[\frac{1}{\sin^2 (\beta + \phi)} - \frac{u^*}{(\Delta \rho) \sin \beta \tan (\beta + \phi)} \right] \quad (16)$$

Combining the terms, we have

$$\tan \gamma = \frac{(u^*/\tau) \tan^2 (\beta + \phi)}{1 + \tan^2 (\beta + \phi) - (u^*/\tau) \tan (\beta + \phi)} \quad (17)$$

where

$$\tau = (\Delta \rho) \sin \beta \quad (18)$$

The quantity τ is the projection of the data spacing in the direction of $\sin \beta$.

If Eq. (17) is solved for u^*/τ , we have

$$u^*/\tau = K(\gamma, \beta + \phi) = \frac{\Delta \tan \gamma [1 + \tan^2 (\beta + \phi)]}{\tan (\beta + \phi) [\tan (\beta + \phi) + \tan \gamma]} \quad (19)$$

Table 3 gives the values of K as a function of angle of detection γ , the sums of the elevation angle β , and one of angle ϕ for the terrain.

The above equation can also be derived from Figure 4 which gives

$$u^*/\tau = K(\gamma, \beta + \phi) = \frac{\Delta}{[\cot (\beta + \phi) - \cot (\beta + \phi + \gamma)]} \quad (20)$$

V. FUNCTION K OF ANGULAR GEOMETRY. If we can keep the ratio u^*/τ constant in Eq. (19) or (20), then the value of $K(\gamma, \beta + \phi)$ will be constant in Table 3. This can be achieved by letting both u^*/\sqrt{R} and τ/\sqrt{R} be constant. First, we will discuss the value of u^*/\sqrt{R} .

Table 1 shows the values of P_F for given threshold to noise ratio T/\sqrt{R} which guarantees the probability of detection. It also lists the values of P_M for both T/\sqrt{R} and the signal to noise ratio, u^*/\sqrt{R} . For example, let us take $T/\sqrt{R} = 1.679$ and $u^*/\sqrt{R} = 3$. We have

$$P_F = 0.0932 \quad (21a)$$

$$P_M = 0.0932 \quad (21b)$$

which are reasonable values for our problem.

The second part is to keep the ratio τ/\sqrt{R} constant. This is related to the scanning scheme given in the next section. From Eq. (18), we have

$$\tau/\sqrt{R} = (\Delta \rho)(\sin \beta)/\sqrt{R} = \text{constant} = L \quad (22)$$

Then the ratios

$$\frac{u^*/\sqrt{R}}{\tau/\sqrt{R}} = \frac{3}{L} = K(\gamma, \beta + \phi) \quad (23)$$

If L is chosen as 4, then one will look at the points for $K = 0.75$ in Table 3. If L is chosen as 1.6, then one will look at the points for $K = 1.875$ in Table 3. For $K = 0.75$, the following set of angles appears

$\beta + \phi$	15°	20°	25°	30°
γ	3.8°	6.7°	10.2°	15.8°

In summary, the above value of γ is guaranteed to be detected for the conditions

TABLE 3. VALUES OF $K(\gamma, \beta + \phi)$ AS FUNCTIONS OF GAMMA AND BETA + PHI

Gamma	5	10	15	20	25	Beta + Phi 30	35	40	45	50
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	3.2857	0.9667	0.4612	0.2724	0.1819	0.1317	0.1011	0.0811	0.0675	0.0578
4	5.1163	1.6605	0.8278	0.5014	0.3405	0.2495	0.1933	0.1562	0.1307	0.1126
6	6.2855	2.1839	1.1270	0.6972	0.4802	0.3557	0.2778	0.2261	0.1902	0.1646
8	7.0986	2.5936	1.3762	0.8668	0.6046	0.4521	0.3558	0.2913	0.2464	0.2142
10	7.6980	2.9238	1.5875	1.0154	0.7164	0.5403	0.4281	0.3527	0.2998	0.2617
12	8.1592	3.1962	1.7694	1.1471	0.8175	0.6214	0.4956	0.4105	0.3506	0.3074
14	8.5258	3.4252	1.9280	1.2649	0.9096	0.6965	0.5589	0.4652	0.3991	0.3514
16	8.8250	3.6210	2.0678	1.3711	0.9941	0.7664	0.6184	0.5172	0.4457	0.3939
18	9.0742	3.7906	2.1922	1.4675	1.0721	0.8316	0.6746	0.5669	0.4905	0.4351
20	9.2856	3.9392	2.3039	1.5557	1.1445	0.8930	0.7279	0.6144	0.5337	0.4751
22	9.4674	4.0709	2.4050	1.6369	1.2120	0.9508	0.7787	0.6600	0.5755	0.5142
24	9.6260	4.1887	2.4972	1.7119	1.2752	1.0055	0.8273	0.7040	0.6161	0.5524
26	9.7658	4.2949	2.5817	1.7818	1.3347	1.0575	0.8738	0.7465	0.6557	0.5898
28	9.8902	4.3913	2.6597	1.8471	1.3910	1.1072	0.9186	0.7877	0.6943	0.6265
30	10.0019	4.4795	2.7321	1.9084	1.4443	1.1547	0.9618	0.8278	0.7321	0.6628
32	10.1030	4.5607	2.7995	1.9662	1.4951	1.2003	1.0037	0.8668	0.7691	0.6986
34	10.1952	4.6357	2.8628	2.0209	1.5436	1.2443	1.0443	0.9050	0.8056	0.7340
36	10.2797	4.7056	2.9223	2.0730	1.5902	1.2868	1.0838	0.9424	0.8416	0.7692
38	10.3577	4.7709	2.9785	2.1226	1.6350	1.3280	1.1224	0.9792	0.8772	0.8042
40	10.4301	4.8322	3.0318	2.1701	1.6782	1.3681	1.1602	1.0154	0.9125	0.8391
42	10.4795	4.8900	3.0826	2.2158	1.7200	1.4071	1.1973	1.0512	0.9476	0.8740
44	10.5608	4.9447	3.1312	2.2597	1.7606	1.4453	1.2338	1.0866	0.9825	0.9090
46	10.6203	4.9968	3.1777	2.3022	1.8002	1.4827	1.2698	1.1218	1.0175	0.9442
48	10.6765	5.0464	3.2225	2.3435	1.8388	1.5195	1.3054	1.1568	1.0524	0.9796
50	10.7299	5.0939	3.2657	2.3835	1.8766	1.5557	1.3407	1.1918	1.0875	1.0154
52	10.7807	5.1396	3.3076	2.4226	1.9136	1.5915	1.3757	1.2267	1.1228	1.0517
54	10.8292	5.1835	3.3482	2.4607	1.9501	1.6269	1.4107	1.2617	1.1584	1.0884
56	10.8758	5.2261	3.3877	2.4981	1.9861	1.6621	1.4456	1.2969	1.1944	1.1258
58	10.9205	5.2673	3.4263	2.5349	2.0217	1.6971	1.4806	1.3323	1.2309	1.1640
60	10.9638	5.3073	3.4641	2.5711	2.0570	1.7321	1.5156	1.3681	1.2679	1.2031

$$u^*/\sqrt{R} = 3, \quad P_F = P_M = 0.0932, \quad \tau/\sqrt{R} = 4$$

and the variable $\beta + \phi$ as listed.

VI. THE SCANNING SCHEME. In order for the value of L to be constant in Eq. (22), one will take the discrete form in Figure 5 as

$$(\Delta\rho)_1 \sin \beta_1 = (\Delta\rho)_2 \sin \beta_2 = \tau = \sqrt{R} L = \text{constant} \quad (23)$$

If b is the height of the mast, then

$$\sin \beta_1 = \frac{b}{\sqrt{b^2 + \rho_1^2}} \quad \sin \beta_2 = \frac{b}{\sqrt{b^2 + \rho_2^2}} \quad (24)$$

Thus, we have in Figure 4

$$\frac{(\Delta\rho)_1}{\sqrt{b^2 + \rho_1^2}} = \frac{(\Delta\rho)_2}{\sqrt{b^2 + \rho_2^2}} = \frac{\tau}{b} = \frac{\sqrt{R}L}{b} \quad (25)$$

The above equation indicates that the spacing of the horizontal projection is proportional to the radial distances from the laser to points on the horizontal plane. For example, let us take $b = 2$, then

$$\frac{(\Delta\rho)_1}{\sqrt{4 + \rho_1^2}} = \frac{(\Delta\rho)_2}{\sqrt{4 + \rho_2^2}} = \frac{\sqrt{R}L}{2}$$

or

$$(\Delta\rho)_i = (\sqrt{4 + \rho_i^2}) \sqrt{R}L/2 \quad (26)$$

For example, given $\sqrt{R} = 0.125$ and $L = 1.6$ or $\sqrt{R} = 0.05$ and $L = 4.0$, in both cases we have $\tau = \sqrt{R}L = 0.20$. Then, from Eq. (26), we have

ρ	2 m	5 m	10 m	20 m
$\Delta\rho$	0.2828 m	0.5353 m	1.1832 m	2.010 m

In this case we may miss a boulder of 0.2 m at 2 m away or a boulder of 2 m at 20 m away.

VII. PROBABILITY OF DETECTION FOR RANDOMLY LOCATED EDGES. In the previous sections it was assumed that range measurements were available from the range finder to the exact vertex of an edge. This is represented by point B in Figures 2 and 3. An edge is defined by a discrete angular change γ . By assuming a range measurement point to fall on the edge vertex, results in a single value for the residual u^* in Eq. (19). This corresponded to a single value for the P_D probability of detection, in which $P_D = 1 - P_M$ with P_M being given by Eq. (6).

In practice, the edges of objects might be randomly distributed and the range measurements, in general, might not fall on an edge vertex. A more realistic approach, therefore, might be to treat the range measurements to be randomly distributed near an edge represented, for example, by points \bar{A} , \bar{B} , and \bar{C} in Figure 6. A random residual \bar{u}^* would result for given edge angular change γ instead of the constant value of u^* assumed previously.

The probability of detection in this case can be calculated using Eq. (6) as a function of the random variable \bar{u}^* :

$$P(\text{Detection}|\bar{u}^*) = 1 - P(\text{Miss}|\bar{u}^*) = 1 - \Phi\left[\frac{T-\bar{u}^*}{\sqrt{R}}\right] \quad (27)$$

From geometric considerations it can be shown that \bar{u}^* will range approximately from $u^*/2$ to u^* in Figure 6. It can be further assumed that the distribution of \bar{u}^* will be uniform over this range of values which corresponds to a purely random distribution of object edges on a given terrain. Equation (27) can then be used to determine the total probability of detection:

$$\begin{aligned} P(\text{Detection}) &= \int_{u^*/2}^{u^*} P(\text{Det}|\bar{u}^*) f(\bar{u}^*) d\bar{u}^* \\ &= \frac{2}{u^*} \int_{u^*/2}^{u^*} P(\text{Det}|\bar{u}^*) d\bar{u}^* \\ &= \frac{2}{u^*} \int_{u^*/2}^{u^*} \left(1 - \Phi\left[\frac{T-\bar{u}^*}{\sqrt{R}}\right]\right) d\bar{u}^* \end{aligned} \quad (28)$$

in which $f(\bar{u}^*)$ equals uniform distribution on $(u^*/2 \text{ to } u^*)$, and u^* is given by Eq. (19).

Equation (28) can be solved as a function of T , \sqrt{R} , and u^* where u^* is a function of γ , $(\beta+\phi)$, and τ as in Eq. (19). As an example, let

$$\begin{aligned} T/\sqrt{R} &= \text{threshold level for detection} \\ &= 2.146; \text{ gives } P_F = 0.0319 \text{ from Table 1} \\ \sqrt{R} &= \sqrt{6}\sigma^2 \text{ where } \sigma = \text{range data noise level} \\ &= 0.1225 \text{ for } \sigma = 0.05 \\ \tau &= 0.20 \end{aligned}$$

The resulting $P(\text{Detection})$ is shown in Table 4 as a function of the angle change γ and $(\beta+\phi)$. Other similar information can readily be generated for other parameter values depending on the actual problem to be solved.

The behavior of the minimum detectable angular change γ can be readily determined from the type of information given in Table 4. For example, by requiring $P(\text{Detection}) = 0.90$ and letting $\beta+\phi = 15$ degrees, results in a minimum detectable value of γ to be about 40 degrees. This compares to a value of γ at 13 degrees for the nonrandom case previously considered where

TABLE 4. TOTAL PROBABILITY OF DETECTION FOR RANDOMLY LOCATED EDGES

Gamma	2	4	6	8	10	Beta + Phi 12	14	16	18	20
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	1.0000	0.9960	0.7526	0.3606	0.1741	0.1004	0.0676	0.0508	0.0412	0.0352
4	1.0000	1.0000	0.9730	0.7563	0.4583	0.2645	0.1636	0.1110	0.0816	0.0641
6	1.0000	1.0000	0.9962	0.9077	0.6794	0.4431	0.2831	0.1891	0.1343	0.1011
8	1.0000	1.0000	0.9993	0.9605	0.8079	0.5895	0.4016	0.2742	0.1941	0.1440
10	1.0000	1.0000	0.9998	0.9812	0.8790	0.6960	0.5055	0.3575	0.2565	0.1901
12	1.0000	1.0000	1.0000	0.9902	0.9199	0.7707	0.5909	0.4338	0.3176	0.2372
14	1.0000	1.0000	1.0000	0.9944	0.9445	0.8230	0.6592	0.5013	0.3754	0.2838
16	1.0000	1.0000	1.0000	0.9967	0.9600	0.8603	0.7133	0.5595	0.4286	0.3286
18	1.0000	1.0000	1.0000	0.9979	0.9703	0.8874	0.7562	0.6094	0.4768	0.3710
20	1.0000	1.0000	1.0000	0.9986	0.9774	0.9076	0.7905	0.6519	0.5202	0.4107
22	1.0000	1.0000	1.0000	0.9990	0.9823	0.9230	0.8181	0.6881	0.5590	0.4476
24	1.0000	1.0000	1.0000	0.9993	0.9859	0.9350	0.8407	0.7191	0.5937	0.4817
26	1.0000	1.0000	1.0000	0.9995	0.9886	0.9444	0.8593	0.7457	0.6246	0.5131
28	1.0000	1.0000	1.0000	0.9996	0.9906	0.9520	0.8748	0.7687	0.6522	0.5421
30	1.0000	1.0000	1.0000	0.9997	0.9922	0.9582	0.8879	0.7887	0.6770	0.5687
32	1.0000	1.0000	1.0000	0.9998	0.9934	0.9632	0.8990	0.8062	0.6993	0.5933
34	1.0000	1.0000	1.0000	0.9998	0.9944	0.9675	0.9085	0.8215	0.7193	0.6160
36	1.0000	1.0000	1.0000	0.9999	0.9952	0.9710	0.9167	0.8351	0.7375	0.6369
38	1.0000	1.0000	1.0000	0.9999	0.9958	0.9741	0.9239	0.8472	0.7539	0.6563
40	1.0000	1.0000	1.0000	0.9999	0.9964	0.9767	0.9302	0.8581	0.7690	0.6743

the range is assumed to be measured directly to the edge vertex.

VIII. CONCLUSION. For an assigned probability of false alarm and miss, one can determine the signal to noise ratio and the threshold to noise ratio. If we use a special scanning scheme such that the spacing of horizontal projections is proportional to the radial distances from the laser to a horizontal plane, then the function K of angular geometry is also constant. The detectible angle γ with certain probability can be found if K and $\beta+\phi$ are known. The angles $\beta+\phi$ are related to the elevation angle β and the terrain slope ϕ .

The following results were discussed in this paper:

1. Ability to detect the change of slopes in a terrain for navigation of vehicles or robotic platforms.
2. Determination of the probability of false alarm and the probability of detection for various signal to noise ratios and for the case of randomly distributed measurements.
3. Determination of required signal to scanning factor ratios for various slopes and slope changes. The signals relate to the probability of detection and the scanning factor influences the size of the obstacles.
4. Computation of the probability for detection of an edge if the reference directions of the laser rays are uniformly distributed over the slopes near an edge point.

APPENDIX

THE SECOND RESIDUAL METHOD

The Second Residual Method, similar to the case case with the Laplacian Method, considers cross-sections of terrain and looks for changes in slope. Here, however, a state estimation and decision process is used to perform the detection. A discrete second order linear time-varying system model is used to estimate ranges and gradients (slopes) from current and previous data. If the difference between a range measurement and range prediction is large enough, a change in slope is indicated and a special estimation scheme is employed.

THE SYSTEM MODEL. A stabilized system model for a terrain has been proposed. The state vector is

$$x_i = [d_i, g_i]^T \quad (A1)$$

where d_i indicates the i -th range and g_i the i -th gradient (or slope). A change of slope is modelled by the presence of an unknown input, u_i , which adds to the gradient component of x_i through the input matrix. The system model is

$$x_{i+1} = F_i x_i + B u_i \quad (A2)$$

where

$$F_i = \begin{bmatrix} b_i & 1-c \\ 0 & q_i \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

where b_i and q_i are the time variant parameters and c is a small non-negative real number necessary for system stability while scanning inward from skylines. The measurement equation is

$$z_{i+1} = H x_{i+1} + v_{i+1} \quad (A3)$$

where $H = [1, 0]$ and v_{i+1} is zero mean Gaussian noise where

$$E\{v_i v_j\} = \begin{cases} 0 & \text{for all } i \neq j \\ \sigma^2 & \text{for all } i = j \end{cases} \quad (A4)$$

Our problem, then, is to detect any small nonzero inputs u_i , since they represent a change in slope.

As in the Laplacian Method, a sufficient statistic is necessary to detect the change of slopes. This sufficient statistic is called the residue, r_{i+2} , as

$$r_{i+2} = z_{i+2} - HF_{i+1}x_{i+1}^* \quad (A5)$$

where x_{i+1}^* is optimal estimate of x_{i+1} given z_{i+1}, z_1, \dots, z_1 . It can be shown that

$$\begin{aligned} E\{r_{i+1}\} &= HF_{i+1}Bu_i \\ &= (1-c)u_i \cong u_i \end{aligned} \quad (A6)$$

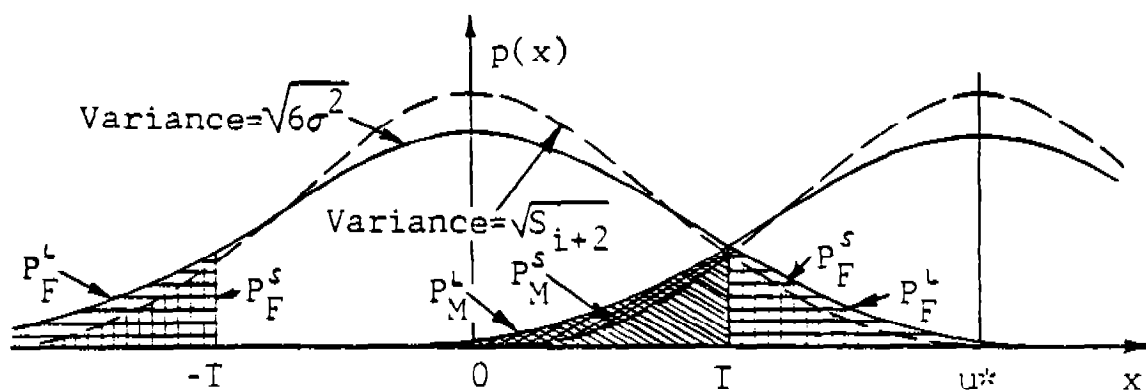


Figure 1. Distributions and Probabilities Comparing the Two Methods.

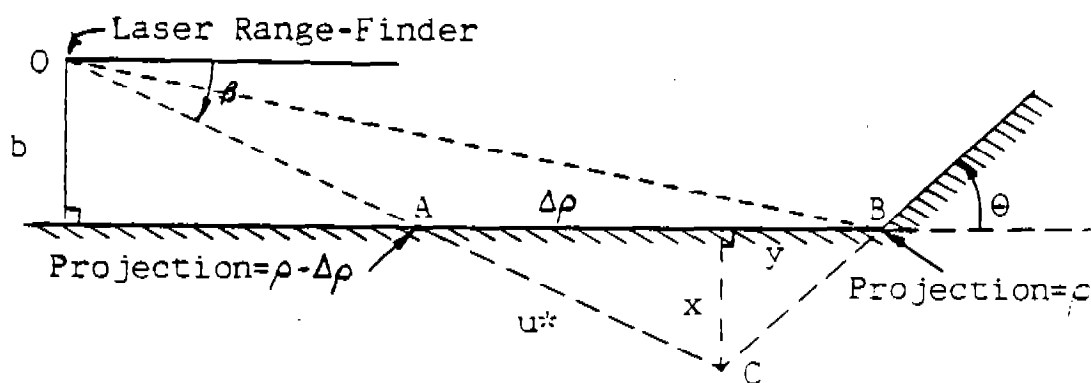


Figure 2. Obstacle Geometry

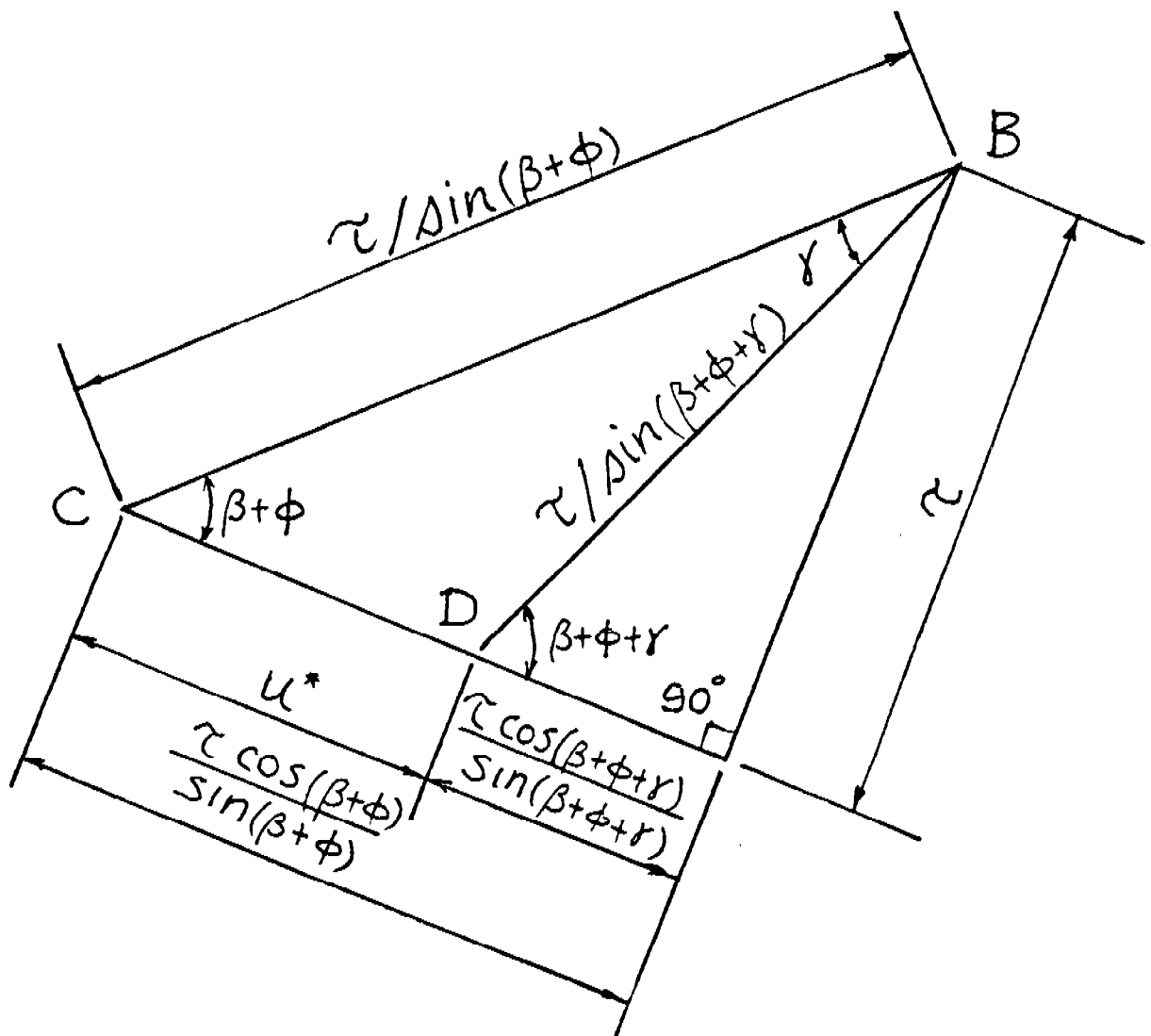


Figure 4. Projection Geometry

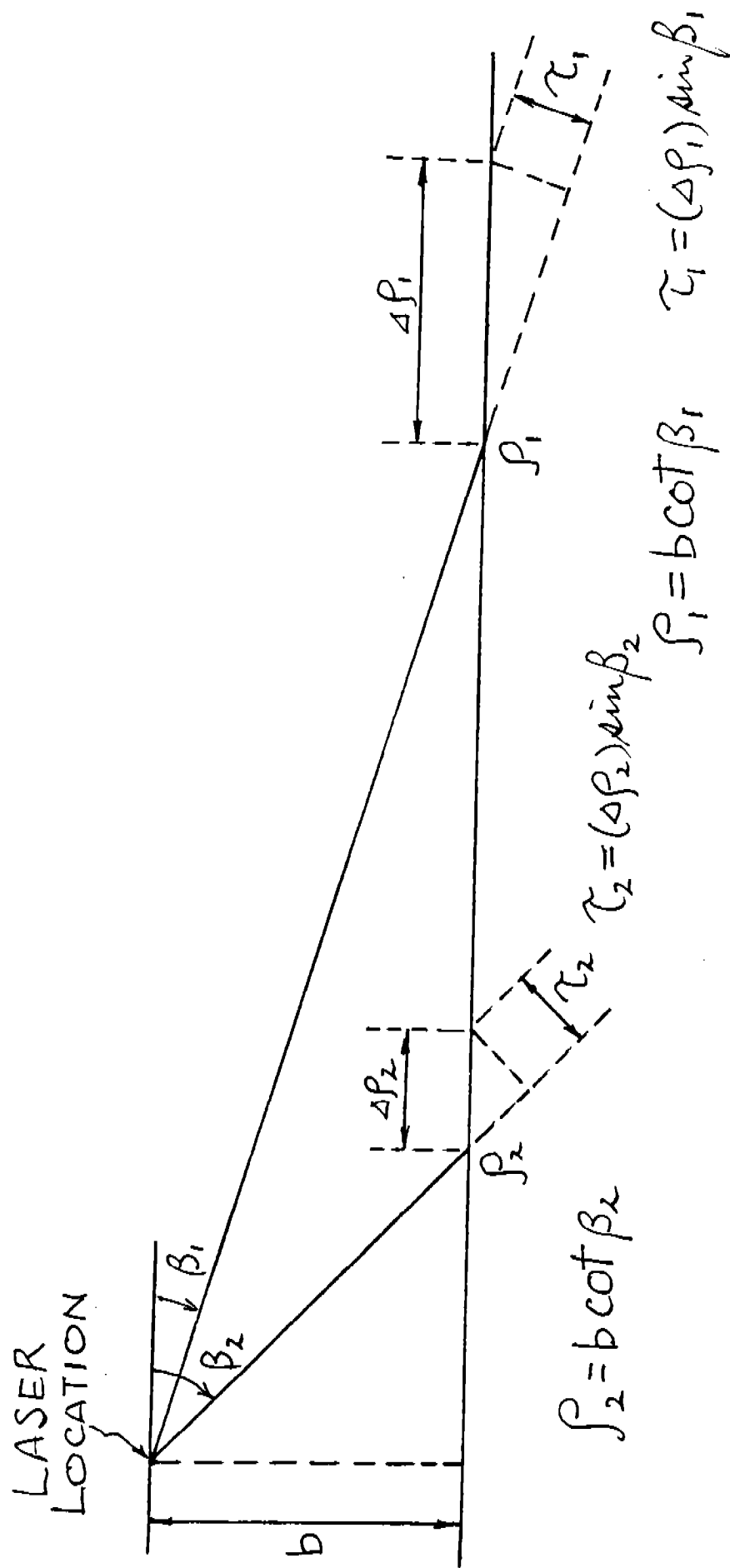


Figure 5. Scanning Factors

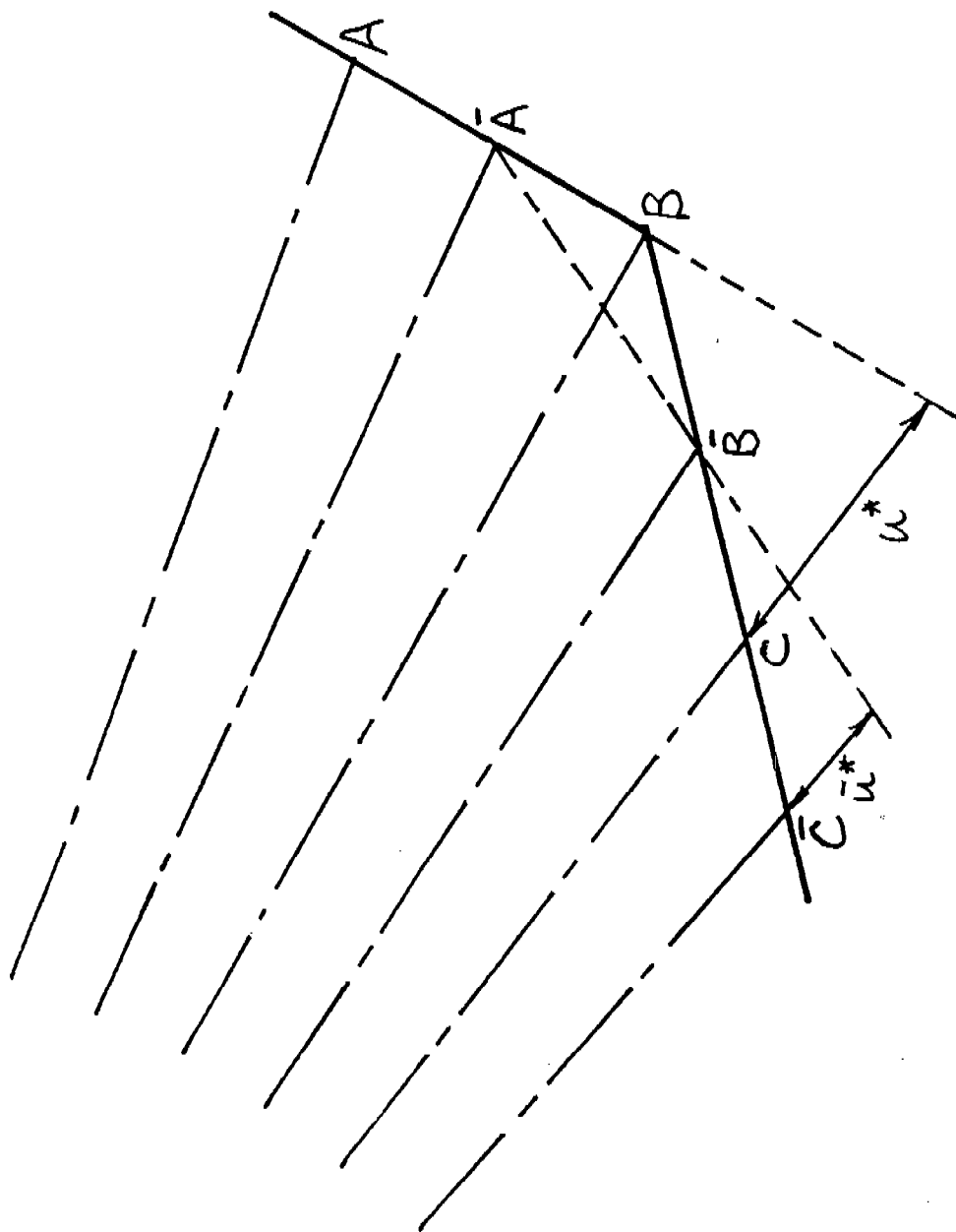


Figure 6. Reference Directions of Laser Rays Near an Edge Point

Identification of Partially Obscured Objects

Charles R. Leake

US Army Concepts Analysis Agency

ATTN: CSCA-RQR

8120 Woodmont Avenue

Bethesda, MD 20814-2797

Abstract. Present computer simulations use an acquisition model that employs the Johnston criteria for identification. This model also provides criteria for different states of acquisition such as detection, recognition, and identification. However, experimental results obtained during tests have shown that it is not always necessary to go through all the stages of acquisition in sequence to achieve identification. Additionally, some unique system characteristics other than size in relation to contrast can be used to acquire objects even when these objects are partially obscured. Other experimentally observed obscuration factors are also discussed which are related to motion and distance for which different stimuli are required for identification. These experimental results are discussed and a mathematical model encompassing these observations is presented.

1. Introduction. This paper is concerned with the image perceived by a viewer using a thermal sight (IR sensor and imaging device). Inasmuch as thermal sights are being used by many of the Army's advanced weapon systems, including tanks and fighting vehicles, as one of their principle means of target acquisition, it becomes important to investigate the human/sight interface in the acquisition process. This paper will concentrate on four factors--atmospheric conditions, thermal image, obscurants, and distance--which have been shown experimentally to influence the image perceived by a human through a thermal sight. A mathematical model of the influence of these factors on the thermal sight and human interface will be presented for possible use in modeling this relationship. A comparison of this model with present techniques will be discussed.

2. Background. Presently equations such as the Beers-Lambert Law or the Bougher-Lambert Law have been modified to account for thermal signature attenuation as a function of temperature contrast between the object and its background. This formula is given in equation 1 below

$$T = T_0 e^{-\delta R} \quad (1)$$

where

T = attenuated temperature in °C

T₀ = object intrinsic temperature contrast in °C

R = Range between object and sensor (km)

δ = Extinction coefficient (Neper per km)

This equation offers a means of determining thermal contrasts as a continuous function of R until some response threshold A_T is reached beyond which the sensor does not function. Acquisition then occurs when $A \geq A_T$ where A is given by equation (2).

$$A = \frac{T - T_B}{T_B} \quad (2)$$

T = Thermal contrast from formula 1.

T_B = Thermal background

Conversely, when $A < A_T$, then acquisition does not occur.

3. Discussion. In reality although differences in temperature between an object and its background affect the target acquisition process, there is more to acquisition than thermal contrasts. For example, there is motion as well as the uniqueness of thermal signatures in relationship to their position relative to the object. To illustrate this point, suppose an object has a temperature T_o and this temperature relative to its background is greater than some threshold t , then this object is discernible by a thermal sensor within the range of the sensor. However, depending on the extent of the background object contrast in relationship to the size of the object, the degree of the acquisition is determined; i.e., detection through identification friend or foe.

This method which is currently used in modeling does not take into account the uniqueness of certain aspects of the thermal signature which are neither related to the temperature of the object nor its size. For example, an exhaust plume from a vehicle which exhibits directionality can in some instances lead to the identification of an object where the object background contrast relationship would indicate that the object should be at the detection phase of the acquisition process. Since the exhaust is separate from the object, unless it is considered in the acquisition process, the relationship between distance and thermal contrasts can be misunderstood.

4. Model.

a. General. In order to include such items as exhaust plumes and other hot areas in thermal signatures, the following model was developed. The three dimensions of the model are atmosphere, image, and distance. Obscurants are an additional dimension which enters into the equation as a temporary condition. The model considers atmosphere and image first, then atmosphere, obscurants and image. It then considers atmosphere, distance and image, and finally atmosphere, distance, obscurants and image.

b. Atmosphere and Image. Let C be a set of characteristic curves on some variable t with G a set of functions on C . For some a we have $n_a, l_a, C_a \subseteq C$ and $g_a \subseteq G$ such that

$$C_{ai} : t \rightarrow \lambda_{ai} \text{ and } g_{ai} : \lambda_{ai} \rightarrow I_{ai} \subseteq I_a$$

n

for $i \in \{1, \dots, n_a\}$. When $\bigcup_{i=1}^{n_a} I_{ai} = I_a$, we have identification.

What this model explains is the manner in which the thermal image is displayed on the screen of a thermal sight. The set C is the set of mappings of the lines on the screen. The image of these mappings is mapped into subportions of the screen by the set of mappings G . For a given atmosphere and object there is an associated image I_a . When the union of the images associated with the subportions of the screen are equal to I_a , the object is then identified. The question of what happens when some but not all of the subportions are missing is addressed in the next section.

c. Atmosphere, Image and Obscurants. Obscurants such as chemical smoke or dust cause a subset of the subportions of the image I_a to be partially obscured. The question then arises as to how well the human can complete the gestalt with missing pieces. This capability happens to be a human trait.

Let p represent obscurant affect with $I_p \subseteq I$. Given an a , we have for $I_{pa} \subseteq I_a$,

$$C_{ai} : t \rightarrow \lambda_{pai} \text{ and}$$

$$g_{ai} : \lambda_{pai} \rightarrow I_{pai}.$$

In this model it is assumed, just as it has been demonstrated in numerous psychological experiments that humans can complete the gestalt. It is also assumed that the human is not just reading into the gestalt in relation to his imagination as would be the case in an ink blot or Roscharch test, but that a true image will be discernible from the partial image I_{pa} . Thus, although it is expected that one would require all the parts, it is possible for a human to estimate all the parts from some of the parts. Thus, identification need not include all the parts in order to occur.

d. Atmosphere, Distance and Image. Clearly C , the set of characteristic curves, will vary with distance d . Thus, for a given distance we have

$$C_{adi} : t \rightarrow \lambda_{adi} \text{ and } g_{adi} : \lambda_{adi} \rightarrow I_{adi} \subseteq I_{ad}$$

Again, if $\bigcup_{i=1}^{n_a} I_{adi} = I_{ad}$, identification has occurred.

In this model the image I_{ad} has been adjusted for distance. This is the case with objects in relation to their background. For example, telephone poles at a distance look smaller than telephone poles that are close up, but everything else in the distance is proportional to the object being identified. This phenomenon is also well documented in elementary psychology texts. It is the adjustment intended for I_{ad} .

e. Atmosphere, Distance, Image, and Obscurants. As discussed in 4c, obscuration can cause a subset of the subportions of the image I_{ad} to be partially obscured. Thus, for a given distance d and an obscurant effect p

$$C_{adi} : t \rightarrow \lambda_{padi} \text{ and } g_{padi} : \lambda_{padi} \rightarrow I_{padi} \subseteq I_{pad}$$

As before, if $\bigcup_{i=1}^{n_a} I_{padi} = I_{pad}$, identification of a partially obscured object has occurred.

This last model is the most general and takes into account the dimensions mentioned in this paper. In all of these models so far motion has not been included and these models represent an object sensed at a fixed distance from the sensor. However, motion does affect one's perception of what is displayed on the screen of a thermal sight and will cause the observer to become aware of something new appearing on the screen as well as continuously changing its position. This change in gestalt included with a unique signature causes objects to be detected and identified simultaneously, which is not predicted from equations (1) and (2). Equations (1) and (2) predict a more gradual transition through the acquisition process, whereas motion and signature allow the observer to accomplish the acquisition process from detection to identification almost instantaneously, and, therefore, at the same apparent range. This has been demonstrated by several studies of which the author is aware (1, 2, and 3).

5. Summary. Target acquisition by a thermal sight as presently modeled through the use of such equations as (1) and (2) requires revisiting. These models predict a gradual transition through the acquisition process. However, actual tests by troops with moving targets indicate that real targets and not ones that have been made to appear like a target are identified at the same time that they are detected. Since the degree of target acquisition is related to the command to fire, this would imply that opening engagements might take place earlier and at greater ranges than would be indicated through the use of equations such as (1) or (2). Inasmuch as the models presented in this paper would be difficult to convert into mathematical formulas, until such equations are developed, a substitute such as actual data found in such studies as 1, 2 and 3 could be converted into look-up tables to determine acquisition as a function of range, atmosphere, obscurants and image to determine when a target was acquired. Moreover, when necessary, interpolation of these tables could be used to provide the needed continuum for use in a computer model. The use of such information could be used to advantage in assisting modelers in discriminating between systems with and without thermal sights as well as for other modeling purposes.

1. USAARENBD, Tank Company Team, Night Fight Test, 1976.
2. USAARENBD, Tank Infrared Elbow (TIRE) FDTE, 1976.
3. USAARENBD, Tank Thermal Sight (TTS), OT11, 1978.

OPTIMUM CONTROL OF FLEXIBLE ROBOT ARMS ON FIXED PATHS.

Sabri Cetinkunt
Wayne J. Book
School of Mechanical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT

Productivity of the industrial robots are directly related to the speed of the task execution. The speed of the robots can be drastically improved by using better control algorithms and reducing the weight of the manipulator.

The speed of a robotic manipulator is constrained by manipulator dynamics and actuator capabilities. Increasing the size of the actuators is not a solution since that will increase the weight of the the overall system leading to a relatively heavier system. The more realistic approach to the problem is to find the optimum control solution for a manipulator to follow a pre-defined path in minimum time, with limited actuator capabilities.

In terms of the dynamic constraints, the weight of the arms may be the most important factor. If a light-weight arm structure is used, actuators will be able to afford higher speeds during the task execution than they would for rigid arm structure. On the other hand using flexible-arms has a major draw-back which is the flexible vibrations, while increasing the speed.

This paper presents the minimum time control solution of a two link flexible arm with actuator constraints. We solved the minimum time problem with no constraints on the flexible modes and show the time improvement due to the use of light-weight arms. The objective is to modify the trajectory, such that flexible vibrations are bounded while changing the solution from the previous one as little as possible. Practical ways of trajectory modifications for flexible arms are discussed.

INTRODUCTION

Today, most trajectory planning algorithms do not consider the dynamics of the manipulators, rather constant and/or piece wise constant accelerations for the overall task are used and an overall maximum allowable speed is set [5,6,7]. However, robotic manipulators are highly nonlinear dynamic systems, so it is expected that affordable accelerations and decelerations and maximum speeds will vary as a function of states. For the traditional schemes to work, the trajectory must be planned for the worst possible case. The capabilities of the system will be used only a small part of the time. Bobrow et.al. [1] first reported that for every point on the path there is an associated maximum allowable speed and maximum affordable acceleration and

This material is based in part on work supported by the National Science Foundation under grant MEA-8303539.

deceleration, and these values can drastically vary from one state to another. Incorporating the manipulator dynamics into the trajectory planning level, they found the minimum time trajectories for different manipulator models [1,2] with limited actuator capabilities moving along pre-defined paths. Shin and McKay [3] solved the same problem independently.

Light-weight manipulators with the same actuator capabilities will be faster. The main problem associated with the light-weight structures is the flexible vibrations. Fig. 1 conceptually shows the performance improvement in terms of increased speed.

In this paper we show the performance improvements due to

1. use of light-weight arms
2. incorporating the manipulator dynamics into trajectory planning level
3. Discuss flexible vibrations during a minimum time trajectory execution and considerations of path modifications such that flexible vibrations will be bounded. This problem is similar in nature to the one raised by Hollerbach [8].

FLEXIBLE MANIPULATOR DYNAMIC MODEL IN JOINT AND PATH VARIABLES

A general dynamic modelling technique for flexible robotic manipulators was developed by Book using recursive Lagrangian-assumed modes method. Homogeneous transformation matrices are used for kinematic relations of the system [4]. A two link flexible robotic manipulator is modelled using that technique (Fig. 2). In the model no actuator dynamics is considered, rather the net torque input to the links is considered as the input variable. No friction at joints nor in the structural vibrations is considered. Flexibility of each link is approximated with one assumed mode for each link. The dynamic model of the manipulator may be expressed in general terms as :

$$[J]_{4 \times 4} \ddot{q} = f(q, \dot{q}) + Q \quad (2-1)$$

where

$$\underline{q}^T: [\theta_1, \theta_2, \delta_1, \delta_2] \quad \text{Joint angle and flexible mode time variables}$$

$$\underline{Q}^T: [\tau_1, \tau_2, 0, 0] \quad \text{Net input torques}$$

$$[J]_{4 \times 4}: \quad \text{Generalized Inertia Matrix, symmetric, positive definite.}$$

$$\underline{f}^T: [f_1, f_2, f_3, f_4] \quad \text{Nonlinear dynamic terms including centrifugal, gravitational, effective spring and Coriolis.}$$

The problem is to find the minimum time trajectories for a given manipulator with limited actuator capabilities moving along a fixed path, with state constraints (bounded flexible vibration constraint). Once the path to be moved along is specified

$$S = S(x, y) \quad (2-2)$$

From inverse kinematic formulation, the corresponding joint angles can be found as

$$\underline{\theta} = \underline{\theta}(s), \quad \underline{\theta}^T = [\theta_1, \theta_2,] \quad (2-3)$$

Similarly, once the speed along the path is known $S(S)$

$$\ddot{\underline{\theta}} = \ddot{\underline{\theta}}(s, \dot{s}) \quad (2-4)$$

and

$$\ddot{\underline{\varepsilon}} = \ddot{\underline{\varepsilon}}(s, \dot{s}, s) \quad (2-5)$$

Knowing the relations (2-3)-(2-5) analytically form or numerically the manipulator dynamics in part can be expressed in path variables.

$$\begin{bmatrix} C_{11}(s, \underline{\varepsilon}) \\ C_{12}(s, \underline{\varepsilon}) \end{bmatrix}_{2 \times 1} \ddot{s} = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} - \begin{bmatrix} C_{21}(s, \dot{s}, \underline{\delta}, \dot{\underline{\delta}}, \vec{e}_t, \vec{e}_n, \rho) \\ C_{22}(s, \dot{s}, \underline{\delta}, \dot{\underline{\delta}}, \vec{e}_t, \vec{e}_n, \rho) \end{bmatrix} \quad (2-6a)$$

$$\begin{bmatrix} \ddot{\delta}_1 \\ \ddot{\delta}_2 \end{bmatrix} = \begin{bmatrix} J_{33} & J_{34} \\ J_{43} & J_{44} \end{bmatrix}^{-1} \begin{bmatrix} f_3 - g_3 + h_1(s, \dot{s}, \vec{e}_t) \\ f_4 - g_4 + h_2(s, \dot{s}, \vec{e}_t) \end{bmatrix} \quad (2-6b)$$

$$\text{where } f_i = f_i(s, \dot{s}, \underline{\delta}, \dot{\underline{\delta}}) \quad (2-7)$$

$$g_i = g_i(s, \dot{s}, \vec{e}_t, \vec{e}_n, \rho) \quad (2-8)$$

$$J_{ij} = J_{ij}(s, \underline{\delta}) \quad (2-9)$$

\vec{e}_t, \vec{e}_n : Unit tangent and normal vectors along the path.

ρ : Curvature of the path at a point.

Notice that flexible modes also affect the position of the end effector, but are not included in the definition of the path. This is mainly due to the fact that we do not have a "direct" control on the flexible vibrations and would like to keep them as small as possible in general.

FORMULATION OF THE NEAR MINIMUM TIME TRAJECTORY PROBLEM FOR FLEXIBLE MANIPULATORS

Using the classical variational calculus principles, the optimum control/programming problem may be stated as:

$$\text{Minimize } J = \int_0^{t_f} dt = \int_{s_0}^{s_f} \frac{ds}{\dot{s}} \quad (3-1)$$

$$s(s_0) = s_0$$

$$\dot{s}(s_f) = \dot{s}_f \quad \text{Initial and final states in path variables.}$$

Subject to :

System dynamics, equations (2-6a) and (2-6b)

Actuator constraints

$$T_{i \min}(\underline{\theta}, \underline{\theta}) \leq T_i \leq T_{i \max}(\underline{\theta}, \underline{\theta}) \quad i = 1, 2 \quad (3-2)$$

Dynamic inequality constraints on flexible modes

$$a_i(t) \leq \delta_i(t) \leq b_i(t) \quad i=1,2 \quad (3-3)$$

The constraints (3-3) naturally arise in flexible structures. If such a constraint is not imposed there is no guarantee on the accuracy of the end point along the path. At first the problem will be solved without considering these constraints. This solution will be used as a nominal solution for the trajectory modification step so that (3-3) are satisfied.

The solution method we use closely follows Bobrow et.al.'s method with some modifications for flexible manipulators. The solution of the above stated optimization problem follows: for any path $S(x,y)$ with given $\dot{S}_0(S_0), \dot{S}_f(S_f)$ to minimize J , \dot{S} should be as large as possible while satisfying the system dynamics and actuator constraints. In order to do so at any state on the path one should use maximum acceleration or deceleration. Then, the problem is reduced to finding the maximum accelerations and decelerations associated with each state of interest. It can be seen from equation (6a) that for each (S, \dot{S})

$$S_d \leq S \leq S_a \quad (3-4)$$

$$S_a = \min \left\{ S_{ai} \right\}$$

$$S_d = \max \left\{ S_{di} \right\}$$

Obviously there may be some range of speeds associated with every point on the path that system can no longer afford to satisfy all conditions (the S range that above inequality is violated). Collection of these ranges defines the forbidden region on (S, \dot{S}) plane. The boundary between allowed and forbidden regions is constant for a given rigid manipulator for a given task. In the case of flexible manipulators, due to the coupling between equations (6a) and (6b) this boundary is also a function of flexible modes, not only (S, \dot{S}) . So, depending on the time history of flexible modes and unpredictable disturbances the boundary will vary. This is not true in the rigid case where the true extremum can be found. At this point the problem is to find when to use maximum accelerations and when maximum decelerations (i.e. to find the switching point(s)). See Fig. 3a-3b.

Finding switching points for flexible manipulators:

1. Integrate $\ddot{S}=\ddot{S}(x,y)$ from final state backward in time until it crosses forbidden region or initial position, using maximum deceleration.
2. Integrate $\ddot{S}(x,y)$ Forward in time with maximum acceleration until the boundary is reached or the two curves crossed each other. If the two curve crossed each other before they enter forbidden region, then find that point. This is the last switching point and terminate the search. If not, then
3. Backup on the forward integrated curve and integrate forward with maximum deceleration until a the trajectory passes tangent to the boundary.
4. Then using the point as new starting point go to step two.

Notice that the last switching point is not the exact switching points, because the flexible modes will not match at this point. That will

cause one to miss the final state somewhat. Also, when searching for the switching points one has to move in a continuous manner in order to keep track of the flexible mode histories accurately. In that sense, the algorithm given at [1] has been modified for flexible robotic manipulators.

SIMULATION RESULTS AND DISCUSSION

The two-link flexible manipulator model for task one (shown in Fig. 4a) was simulated for the two different cases in order to show the performance improvement achieved due to light-weight system. In both cases actuators have same capabilities. It is found that weight reduction by a factor of 2 results in approximately 60 % time improvements (Fig. 5a and 6a). This improvement, of course, slightly varies depending on the task. Joint actuator histories are shown in Fig. 5b-6c and flexible mode responses are shown in Fig. 5c-6d.

Task 2 (Shown in Fig. 4b) simulated for light-weight manipulator and results are shown Fig 7 a-d. The final trajectory is shown in heavy lines. One interesting point in this simulation is the fact that as soon as the manipulator end point enters the curvature the system must accelerate along the path in order to obey the constraints. In Fig. 5a the curve ab shows that right before the curvature the system is able to afford deceleration (aa' curve), but as end point enters the curvature, then the sudden appearance of a normal acceleration term in the dynamics of the system makes the difference. The other point in the case of flexible arms is that at the last switching point flexible modes are not same, since they have different histories. This will cause error in the final state reached. See Fig. 6a, 7a. The last switching point needs to be varied from the original result of the above algorithm. This can be done on trial and error basis at the trajectory planning level.

5. CONCLUSION AND FURTHER WORK

In this paper we showed ways to improve performance and productivity of Robotic manipulators With Flexible arms. One way was to use light-weight structures and the other was to incorporate the dynamics of manipulators in to trajectory planning level and make optimum utilization of given manipulator. This method can be used with any path. Application of the method requires manipulator model, Geometric path in work space, and actuator capabilities. Obviously as trajectory gets closer to the forbidden region boundary system capabilities are being used to the limits and any disturbance or uncertainty can easily put the system into forbidden region. The situation is more dramatic for flexible manipulators. While this analysis is nice in terms of knowing the ultimate capabilities, in practice there will be a safety factor that will require to keep the optimal trajectory away from the forbidden region certain amount. Research is in progress on the Optimum modification of the trajectories found by above described method so that inequality constraints on the flexible modes will be satisfied.

REFERENCES

1. Bobrow, J.E., Dubowsky, S., Gibson, J.S. "On the Optimal Control of Robotic Manipulators with Actuator Constraints" Proc. of 1983 ACC, San Francisco, CA June 1983, pp 782-787

2. Dubowsky, S., Shiller, Z. "Optimal Dynamic Trajectories For Robotic Manipulators" Fifth CISM-IFTOMM Symposium On The Theory And Practice Of Robotic Manipulators Preprints, June 26-29 1984, Udine, Italy, pp 96-103.
3. Shin, K.G., McKay, M.D. "Minimum-Time Control Of Robotic Manipulators With Geometric Path Constraints". IEEE Trans. on Automatic Control, Vol AC-30 No.6, June 1985 pp 531-541.
4. Book, W.J. "Recursive Lagrangian Dynamics of Flexible Manipulators" The International Journal of Robotic Research, MIT Press, V.3, N.3 pp. 87-101, Fall, 1984.
5. Kahn, M.E., Roth, B. "The Near-Minimum time Control of Open-loop Articulated Kinematic Chains" Journal of Dynamic Systems, Measurement, and Control, ASME Trans., Vol. 93, No. 3, Sept 1971, pp 141-171.
6. Luh, J.Y.S., Lin, C.S. "Optimum Path Planning For Mechanical Manipulators" Journal of Dyn. Syst. and Measurement and Control, ASME Trans., Vol. 102, No. 2, June 1981, pp 142-151.
7. Luh, J.Y.S., Walker, M.W., "Minimum-time Along the Path for a Mechanical Manipulator", Proc. of IEEE Conf. on Decision and Control, Dec. 1977, New Orleans, LA, pp 755-759.
8. Hollerbach, J.M. "Dynamic Scalling Of Manipulator Trajectories" Proc. of ACC, June 1983, San Francisco, CA.

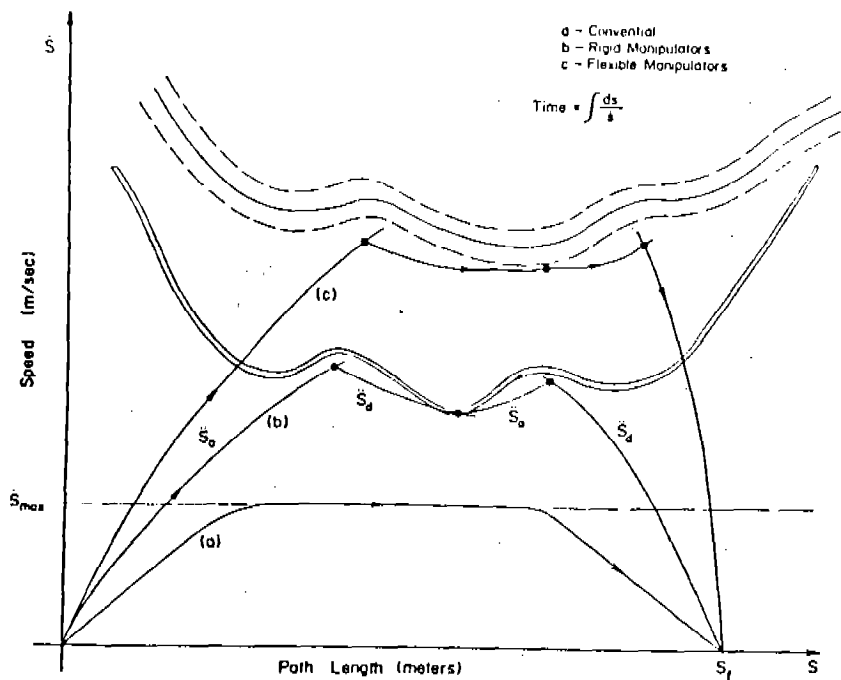


Fig.1 Different trajectory plans.

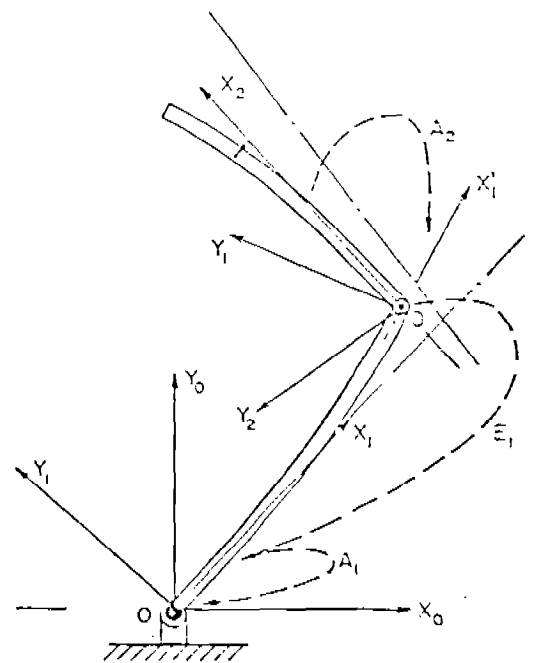


Fig.2 Manipulator Model.

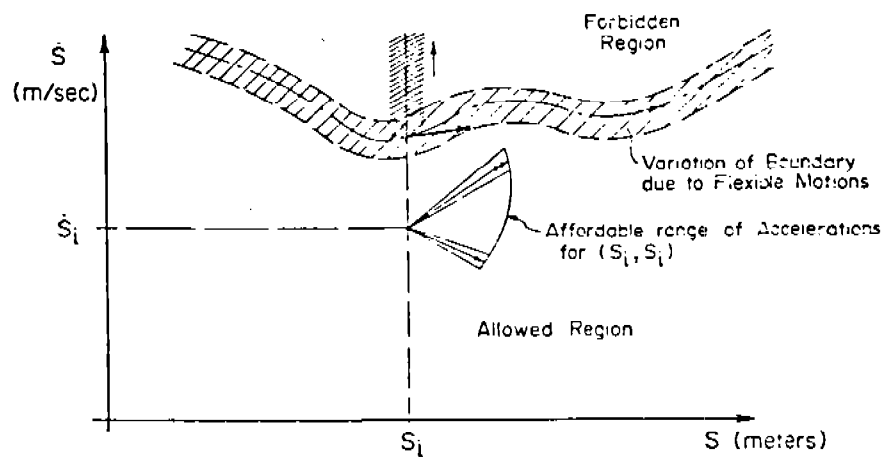


Fig. 3.a (S, \dot{S}) Plane

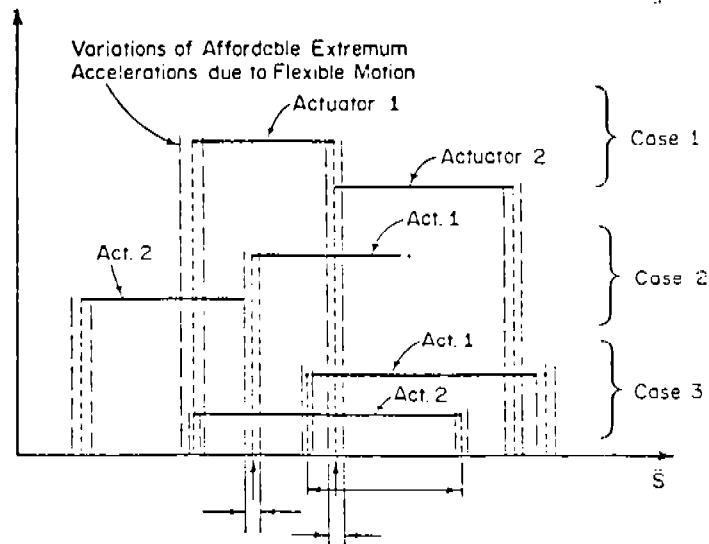


Fig. 3.b Different possible cases during a task.

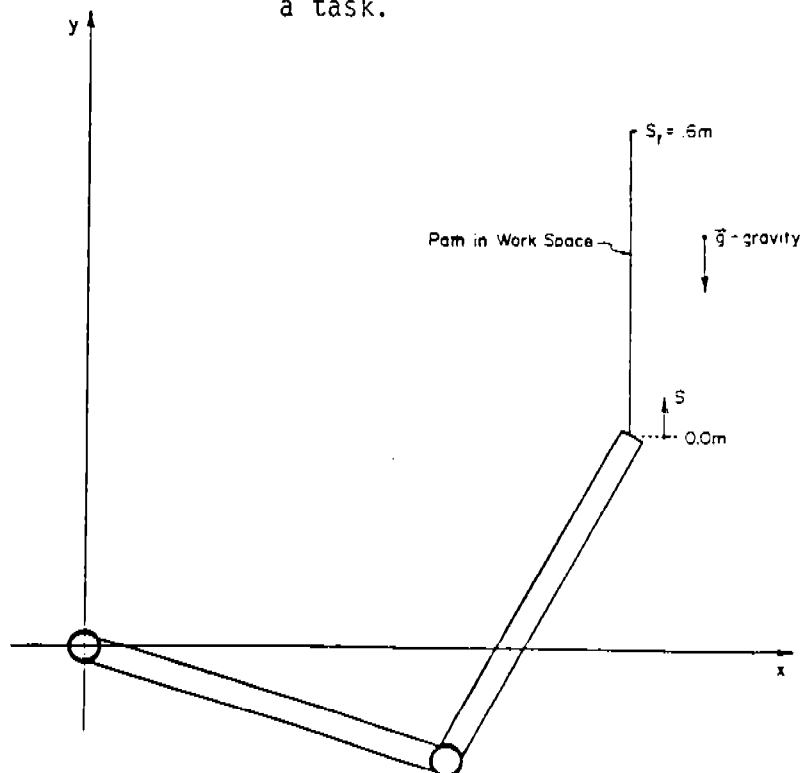


Fig.4.a Task 1 in (x,y) plane.

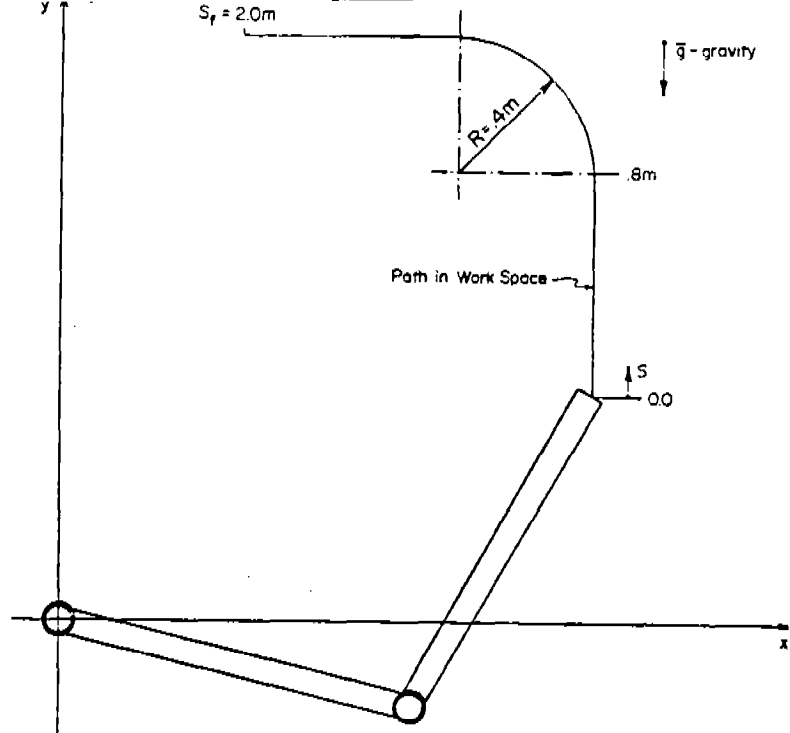


Fig. 4b Task 2 in (x,y) plane.

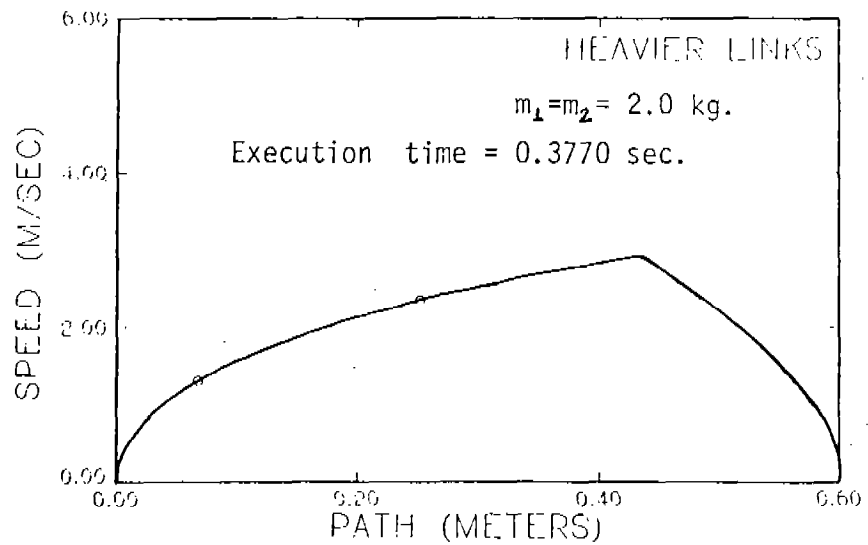


Fig. 5.a Trajectory for Path 1 of heavy links.

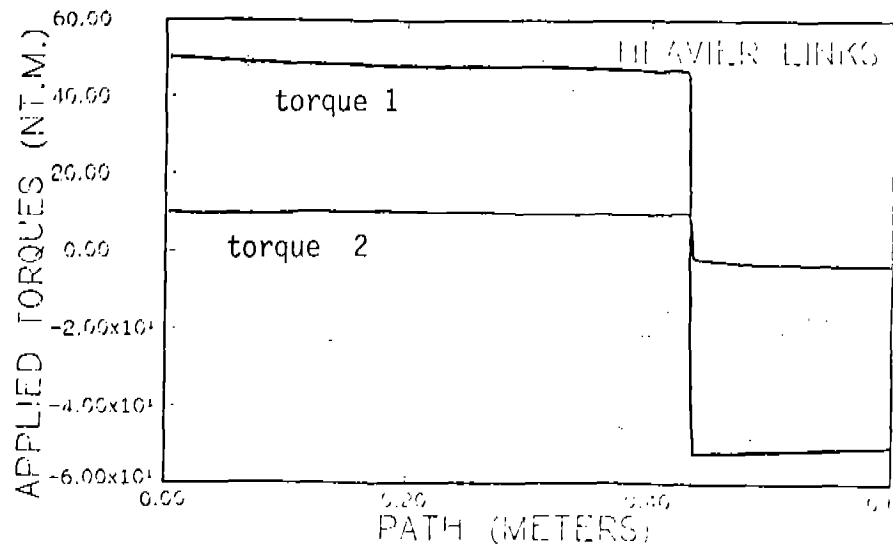


Fig. 5.b Torque histories for path 1.

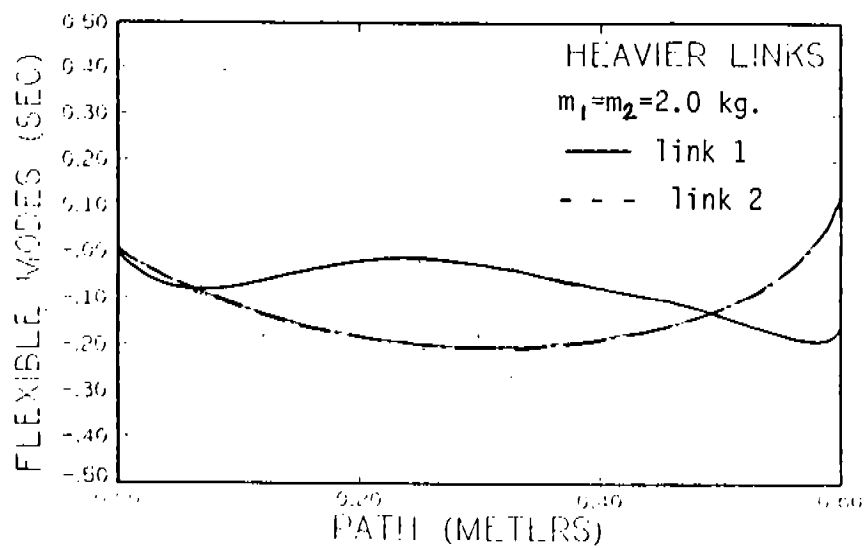


Fig. 5.c Flexible modes time variables

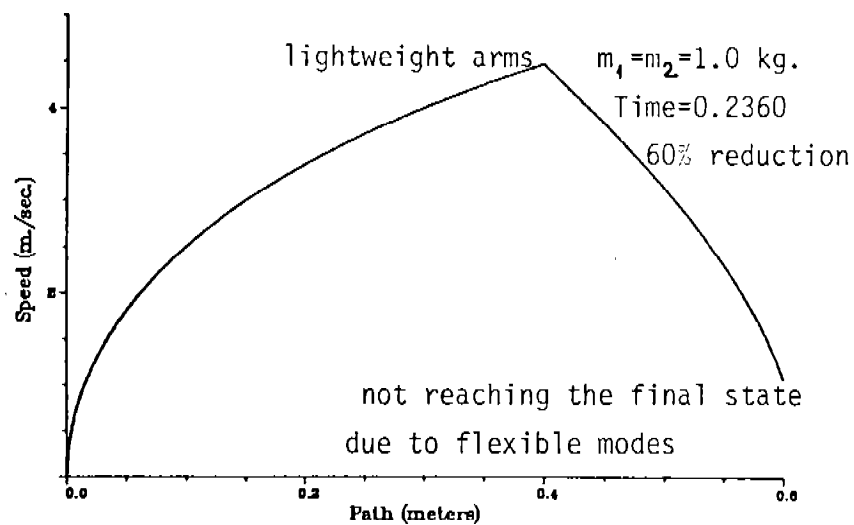


Fig. 6.a Trajectory of lightweight arms.

Find Switch P.

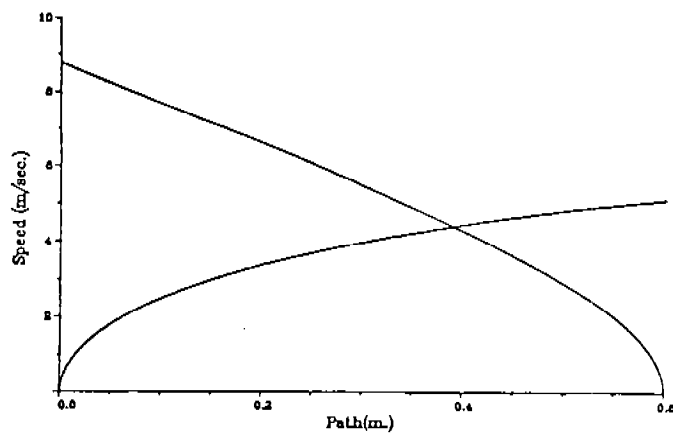


Fig. 6.b Finding the switching point

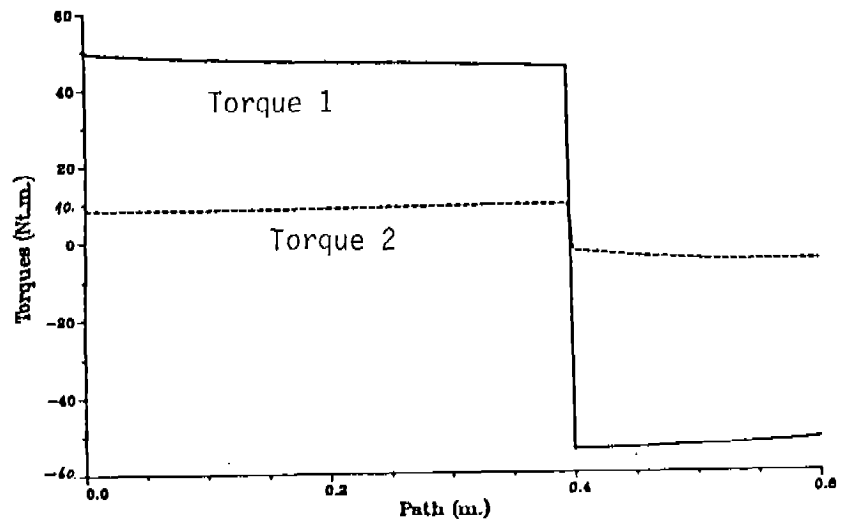


Fig. 6.c Torque histories of lightweight arms along path 1.

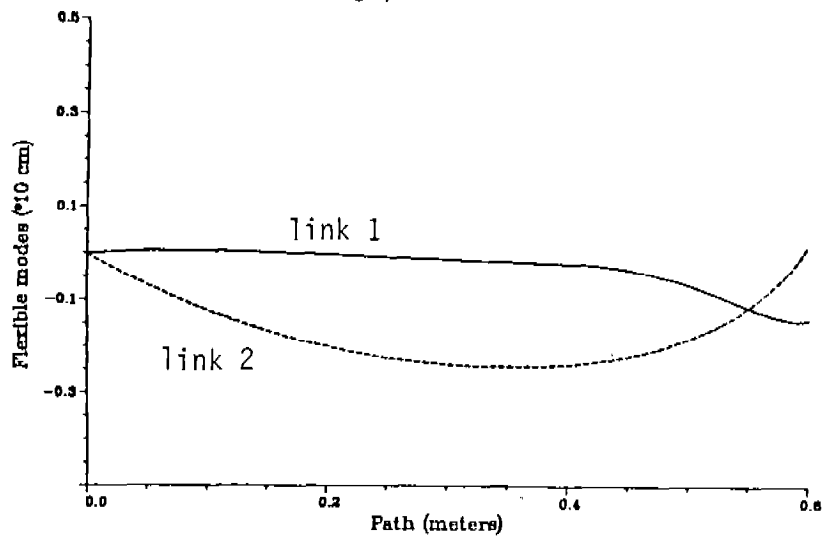


Fig. 6d. Flexible modes along path 1.

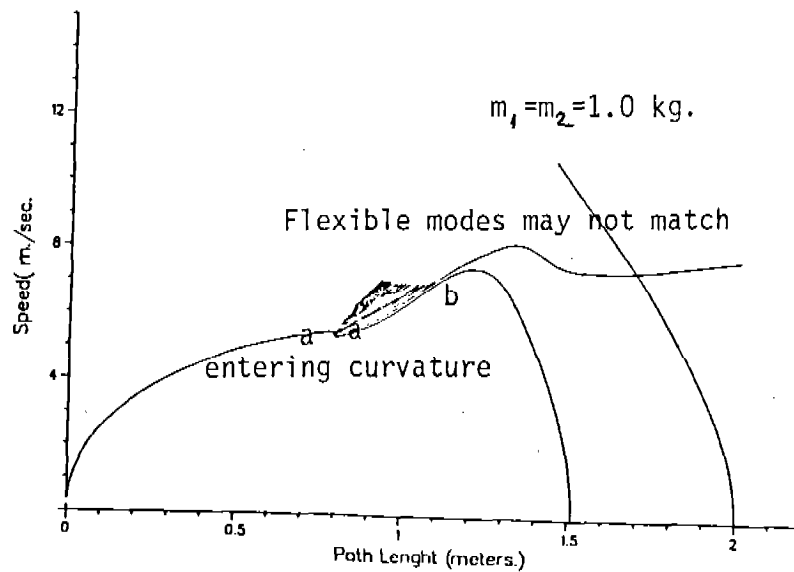


Fig. 7.a Finding the switching points for path 2.

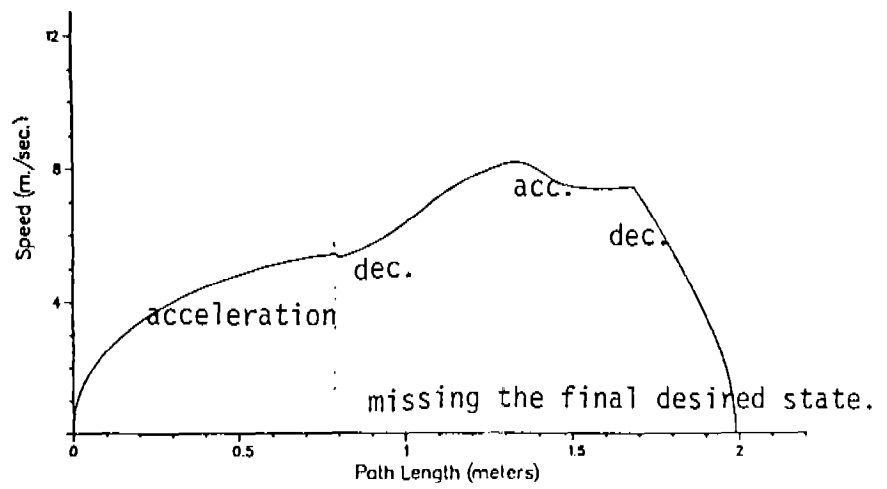


Fig. 7b. Trajectory for path 2.

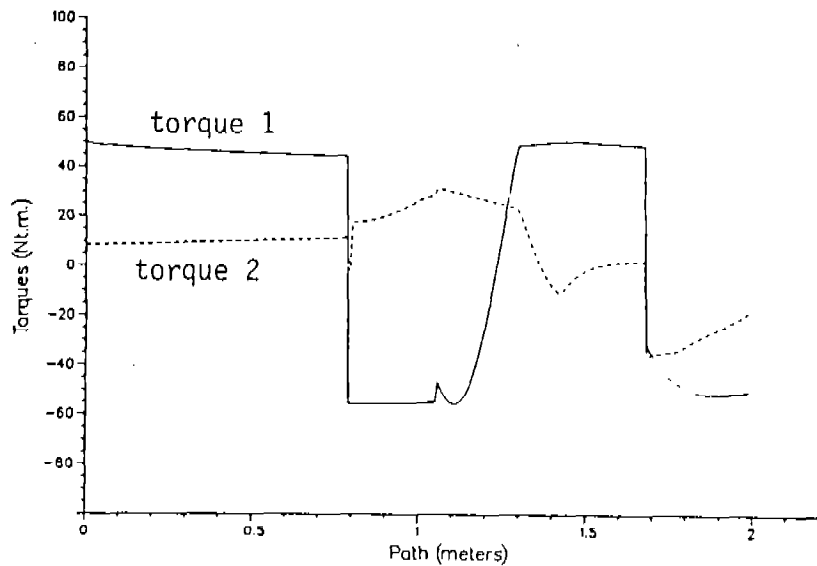


Fig. 7c . Torque histories along path 2

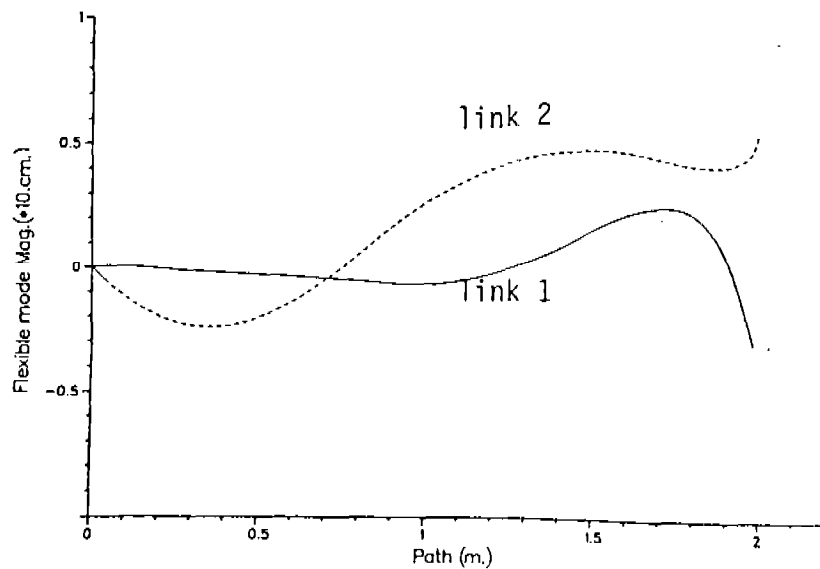


Fig. 7.d Flexible modes along path2.

DYNAMICS OF FLEXIBLE MECHANICAL SYSTEMS

Wan S. Yoo and Edward J. Haug
Center for Computer Aided Design
and
Department of Mechanical Engineering
College of Engineering
The University of Iowa
Iowa City, Iowa 52242

ABSTRACT. A finite element based method is developed and applied for geometrically nonlinear dynamic analysis of flexible systems. Vibration and static correction modes are used to account for linear elastic deformation of components. Boundary conditions for vibration and static correction mode analysis are defined by kinematic constraints between components of a system. Constraint equations between flexible bodies are derived and a Lagrange multiplier formulation is used to generate the coupled large displacement-small deformation equations of motion. A standard, lumped mass finite element structural analysis code is used to generate deformation modes and deformable body mass and stiffness information. An intermediate-processor is used to calculate time-independent terms in the equations of motion and to generate input data for a large scale dynamic analysis code that includes coupled effects of geometric nonlinearity and elastic deformation. Two examples are presented and the effects of deformation mode selection on dynamic prediction are analyzed.

I. INTRODUCTION. Developments in multibody dynamics have come from two principle sources; mechanism dynamics and spacecraft dynamics. Early developments in the field of flexible mechanisms are typified by Refs. 1-5. While numerous technical differences exist between the methods used in these studies, the pervasive assumption made is that large amplitude gross motion dynamics can be uncoupled from small amplitude elastodynamics of the system.

Attempts have been made to account for coupling between gross motion and elastodynamics using continuum models and deriving special purpose equations of motion. A mechanism example that typifies this approach is presented in Ref. 6. A related approach to spacecraft dynamics may be found in Refs. 7-9.

Essential limitations associated with the assumptions of uncoupling gross motion and vibration, using elastic continuum models, has led workers in both fields of flexible mechanism dynamics and spacecraft dynamics to formulations that employ finite element based techniques to represent flexibility and coupling these effects with gross motion dynamics. Bodley and co-workers developed a computer program called DISCOS [10] for spacecraft dynamics and control, including flexibility

effects through finite element modal analysis. Related developments in spacecraft structural dynamics with open loops may be found in Refs. 11 and 12. A substantial extension of these methods, using relative joint coordinates and accounting for closed loops, has recently been presented by Keat [13].

Finite element based, fully coupled dynamic formulations for machine dynamics began in the late 1970's and are reflected by the early papers of Refs. 14 and 15. Song [14] presented a general formulation for planar system dynamics that incorporated flexible finite elements into a general purpose, rigid body mechanism dynamics code. On the other hand, Sunada and Dubowsky [15] used a lumped mass finite element model of structural components and developed the equations of motion for selected machines. Their work provided substantial insight into use of finite element data in generating the equations of motion. Shabana and Wehage [16] extended the method of Ref. 14, to include vibration modes to model flexible bodies in a general purpose dynamics code. The key contribution from this work was clear demonstration that flexibility effects can be included in a general purpose spatial dynamics code. The limitation of the approach of Ref. 16 is that flexible bodies must be made up of collections of finite elements that are imbedded in the analysis code.

The purpose of this paper is to present and illustrate use of a finite element based numerical method for dynamic analysis of mechanical systems that contain complex-shaped, flexible bodies. To achieve this goal, the following approach is employed;

- (1) Vibration and static correction deformation modes are used to define kinematically admissible deformation fields in elastic components.
- (2) Constraint equations between flexible bodies are derived.
- (3) The equations of motion are derived, using lumped mass finite element approximations.
- (4) A standard finite element structural analysis code is used for vibration and static correction mode analysis of each flexible body. A preprocessor calculates time-independent terms in the equations of motion and generates input data for geometrically nonlinear dynamic analysis of flexible systems.

II. GENERALIZED COORDINATES AND KINETIC ENERGY. In order to specify the state of a body, it is necessary to define a set of generalized coordinates that uniquely locate every point in the body in space. The XYZ reference frame shown in Fig. 1 represents an inertial reference frame and the xyz frame is a reference frame for a typical body, which need not be fixed to the body as it deforms. The xyz frame is defined to be fixed to the body in its undeformed state; i.e., points on the body can be defined relative to the xyz frame, prior to deformation. If the distributed mass around node i is replaced by a point mass at that node, which is called lumped mass, that mass has no rotary inertia.

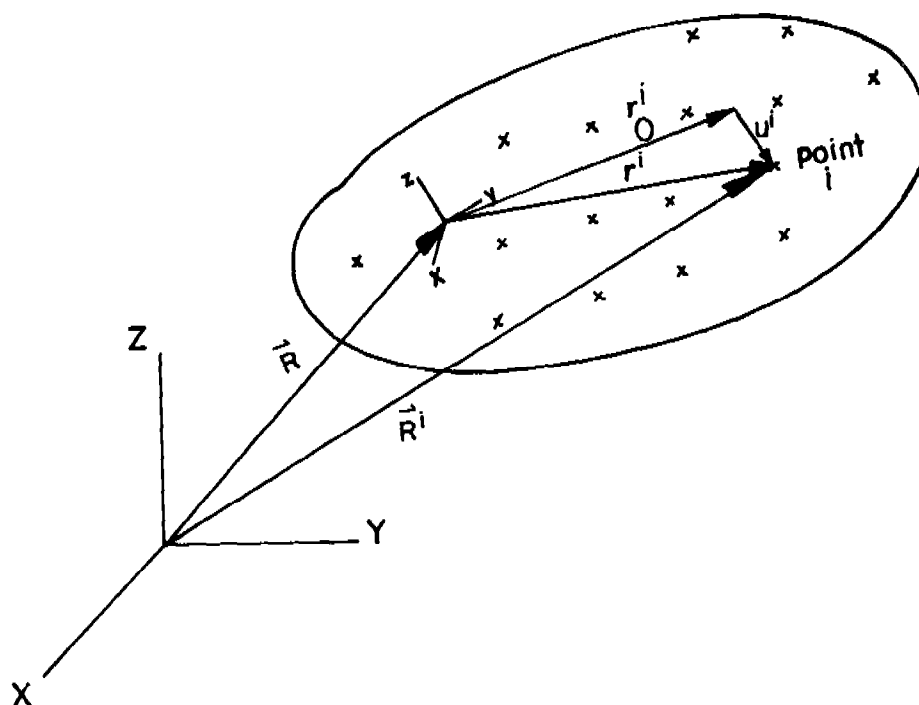


Figure 1. Reference Generalized Coordinates

Consider node point i on the body, defined by the vector r_0^i in the xyz frame in the undeformed state. Let P^i be a projection matrix that extracts the xyz coordinates of the displacement u^i of point i due to deformation; i.e.,

$$u^i = P^i u \quad (1)$$

where u is the nodal displacement vector for the entire body, which contains both nodal displacements and rotations. After deformation, point i is located in the xyz frame by

$$r^i = r_0^i + u^i = r_0^i + P^i u \quad (2)$$

where r^i and r_0^i are vectors from the origin of the xyz frame to point i in the deformed and undeformed states respectively. Global XYZ coordinates of point i are thus

$$R^i = R + A(r_0^i + P^i u) \quad (3)$$

where A is the transformation matrix from the xyz frame to the XYZ frame.

In order to uniquely characterize the state of the body, in terms of generalized coordinates of the reference frame and deformation mode coordinates, one must exclude modes corresponding zero frequencies. Otherwise, rigid body motion is represented by both R and the modal coordinates and uniqueness of state as a function of generalized coordinates is lost.

In terms of deformation mode coordinates, one can calculate the velocity of point i from Eq. 3 as

$$\dot{R}^i = \dot{R} + \dot{A}(r_0^i + p^i \psi a) + A p^i \dot{\psi} \dot{a} \quad (4)$$

where ψ and a are modal matrix and modal coordinate vector, respectively. Using Euler parameters [17] (See Appendix) as rotational generalized coordinates of the xyz frame, relative to the XYZ frame, the absolute velocity becomes

$$\begin{aligned} \dot{R}^i &= \dot{R} - 2E r^{+i} \dot{p} + A p^i \dot{\psi} \dot{a} \\ &= \dot{R} - 2A \tilde{r}^i G \dot{p} + A p^i \dot{\psi} \dot{a} \end{aligned} \quad (5)$$

where p is the (4×1) vector of Euler parameters, r^i is any (3×1) vector defined in the xyz frame, r^{+i} and \tilde{r}^i are (4×4) and (3×3) matrices composed of the elements of r^i , and E and G are (3×4) matrices defined by the elements of p (see Appendix for details). Summing the kinetic energy of masses m^i lumped at nodes i over the total number N of nodes yields the kinetic energy of the body [18] as

$$T = \frac{1}{2} \begin{bmatrix} \dot{R} \\ \dot{p} \\ \dot{a} \end{bmatrix}^T M \begin{bmatrix} \dot{R} \\ \dot{p} \\ \dot{a} \end{bmatrix} \quad (6)$$

where

$$M = \begin{bmatrix} \left(\sum_{i=1}^N m_i \right) I_3 & -2E \sum_{i=1}^N m_i \dot{r}^i & A \sum_{i=1}^N m_i p^i \psi \\ & 4G^T \left(\sum_{i=1}^N m_i \tilde{r}^i \tilde{r}^{iT} \right) G & -2 \sum_{i=1}^N m_i G^T \tilde{r}^i \tilde{r}^{iT} p^i \psi \\ & & + 2\rho \sum_{i=1}^N m_i r^i \tilde{r}^{iT} p^i \psi \\ \text{symmetric} & & \sum_{i=1}^N m_i (p^i \psi)^T (p^i \psi) \end{bmatrix}$$

Note that vector r^i appearing in the mass matrix depends on both reference and modal coordinates, so it is not constant. Evaluation of M at each time step of a simulation requires expansion of terms that involve generalized coordinates. An efficient method of carrying out the expansion and evaluating M is presented in Refs. 18 and 19.

III. STRAIN ENERGY. The strain energy of a deformable body is calculated in the finite element code as

$$U = \frac{1}{2} u^T K u \quad (7)$$

where K is the structural stiffness matrix. Using modal matrix and coordinates, Eq. 7 becomes

$$U = \frac{1}{2} a^T \Psi^T K \Psi a = \frac{1}{2} a^T K_{aa} a \quad (8)$$

where

$$K_{aa} = \Psi^T K \Psi$$

IV. GENERALIZED FORCE. Let F^i be an external force acting on node i of a body. The virtual work of all such forces acting on the body can be written as

$$\delta W = \sum_{i=1}^N F^i{}^T \delta R^i \quad (9)$$

The total differential of Eq. 3 can be used to calculate the virtual displacement of node i as

$$\delta R^i = \delta R - 2A \tilde{r}^i G \delta p + A p^i \Psi \delta a \quad (10)$$

Equation 9 can now be written as

$$\delta W = [Q_R^T \quad Q_p^T \quad Q_a^T] \begin{bmatrix} \delta R \\ \delta p \\ \delta a \end{bmatrix} \quad (11)$$

where

$$\left. \begin{aligned} Q_R &= \sum_{i=1}^N F^i \\ Q_p &= \sum_{i=1}^N 2G^T \tilde{r}^i A^T F^i \\ Q_a &= \sum_{i=1}^N (A p^i \psi)^T F^i \end{aligned} \right\} \quad (12)$$

Forces due to spring, damper, and actuator elements are included in the generalized force vector Q .

V. EQUATIONS OF MOTION OF A FLEXIBLE BODY. The equations of motion of a flexible body with no kinematic constraints can be written as

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}} \right)^T - \left(\frac{\partial T}{\partial q} \right)^T + \left(\frac{\partial U}{\partial q} \right)^T + J^T \lambda = Q \quad (13)$$

where T is kinetic energy of the body, $q = [R^T, p^T, a^T]^T$, U is total strain energy of the body, Q is generalized force acting on the body, J is the Jacobian matrix of the Euler-parameter constraint equation $p^T p = 1$ [17], and λ is a Lagrange multiplier associated with that constraint.

Expanding the kinetic energy expression of Eq. 6 and using the strain energy expression of Eq. 8, the equations of motion of the body are formulated [18]. In these equations of motion, many calculations are required to evaluate all terms. If terms are partitioned into time-dependent and time-independent subsets, the amount of calculation is substantially reduced [18]. Time-independent terms are precalculated in a subroutine that takes input from an established finite element structural analysis code. The resulting values are then read into the geometrically nonlinear dynamics code to form the equations of motion (Eq. 13) of the body. For a detailed development of the flexible body equations of motion and a discussion of methods used to generate deformation modes using a finite element code, the reader is referred to Refs. 18 and 19.

VI. EQUATIONS OF MOTION OF A CONSTRAINED SYSTEM. System equations of motion can be obtained by combining equations of motion of each body and equations of constraint between bodies, using Lagranges equations of motion, as

$$\frac{d}{dt} \left(\frac{\partial \dot{T}}{\partial \dot{q}} \right)^T - \left(\frac{\partial T}{\partial q} \right)^T + \left(\frac{\partial U}{\partial q} \right)^T + J^T \bar{\lambda} = \bar{Q} \quad (14)$$

where T is total kinetic energy of the system, \bar{q} is the complete set of generalized coordinates for the system, U is the strain energy of the system, \bar{Q} is the complete vector of generalized forces, J is the Jacobian matrix of the kinematic constraint equations (including $p^T p = 1$ for each body), and $\bar{\lambda}$ is the complete vector of Lagrange multipliers. This system of equations can be systematically assembled from Eq. 13 for each body and the constraint equations derived in Section 7.

VII. CONSTRAINT FORMULATION. Kinematic constraints between flexible bodies must account for deformation of bodies that are connected. Otherwise, derivation of constraint equations proceeds as in the case of rigid body systems. In this paper, the equations for spherical, universal, and revolute joints are formulated for rigid and flexible bodies.

(a) Spherical Joint (SPHR). A spherical joint at point P between two adjacent bodies i and j is shown in Fig. 2. A vector equation that requires point P to be common to both bodies is

$$R_i + A_i s_i' - R_j - A_j s_j' = 0 \quad (15)$$

where s_i' and s_j' are vectors from the origins of the ith and jth local reference frames to point P. Let s_{i0}' and s_{j0}' be values of s_i' and s_j' in the undeformed state, measured in the respective body reference frames.

If both bodies are flexible, s_i' and s_j' depend on elastic deformation and must be evaluated at each deformed state. If modal coordinates are employed to represent elastic deformation, Eq. 15 can be written as

$$R_i + A_i [s_{i0}' + (p^k \psi_a)_i] - R_j - A_j [s_{j0}' + (p^k \psi_a)_j] = 0 \quad (16)$$

where A_i and A_j are transformation matrices from the ith and jth local reference frames to the global frame and k is the number of the node at which the spherical joint is located. This joint generates three algebraic constraint equations.

(b) Universal Joint (UNIV). A universal joint between adjacent bodies i and j is shown in Fig. 3. Three points, P in both bodies, I in

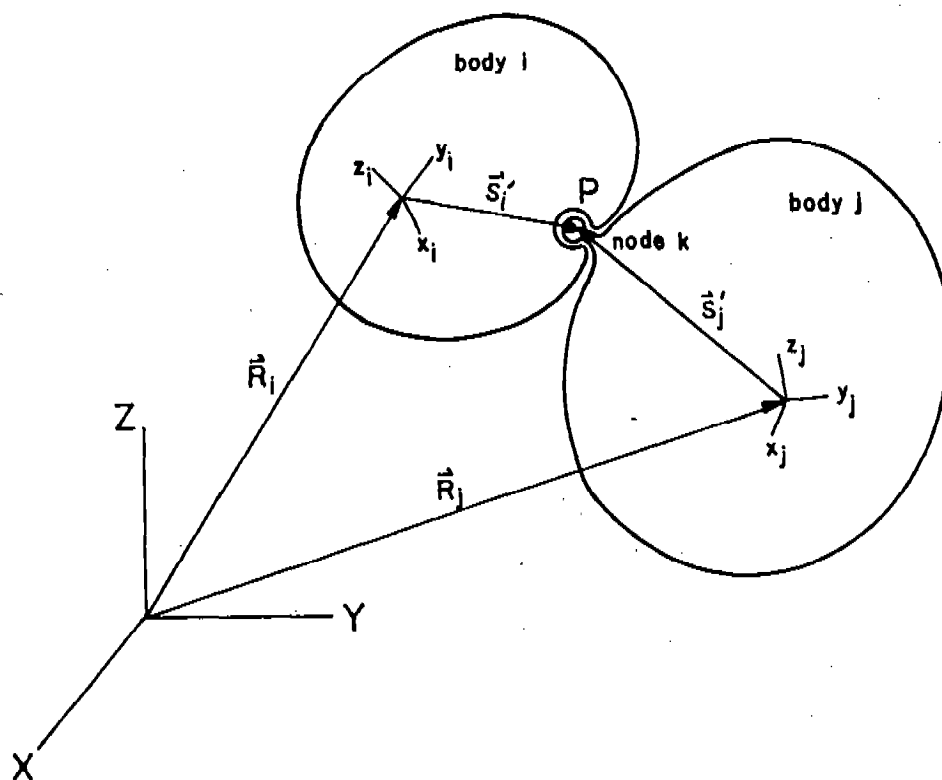


Figure 2. Spherical Joint

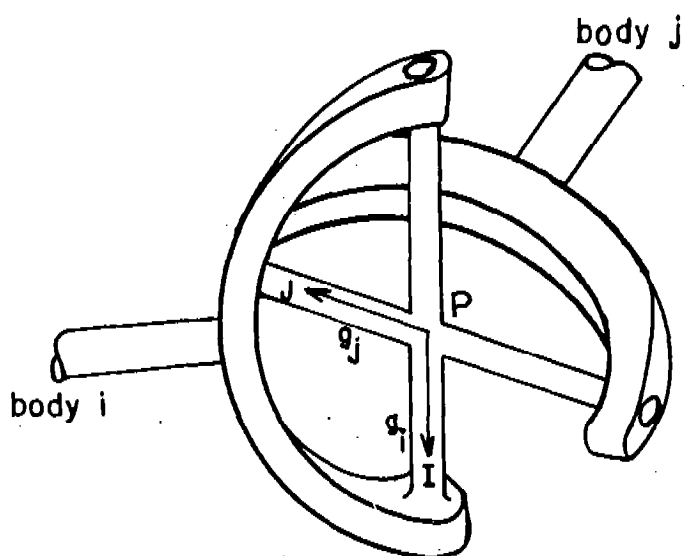


Figure 3. Universal Joint

body i , and J in body j are chosen to define a universal joint. Point P must satisfy the spherical joint equation of Eq. 16. The constraint that requires vectors g_i and g_j , defined in Fig. 3, to be perpendicular is that their scalar (dot) product is zero, denoted as the DOT1 constraint,

$$g_i^T g_j = 0 \quad (17)$$

where g_i and g_j are vectors from point P to points I and J in the global frame.

When both bodies are flexible, elastic deformation should be considered. To include the effect of rotational deformation, define body-fixed $\xi_i \eta_i \zeta_i$ and $\xi_j \eta_j \zeta_j$ frames at point P in each body. It is assumed that the structure near the joint is sufficiently stiff so that vectors g_i'' and g_j'' defined in the $\xi_i \eta_i \zeta_i$ and $\xi_j \eta_j \zeta_j$ frames remain constant. Then, Eq. 17 can be written as

$$(A_i B_i g_i'')^T (A_j B_j g_j'') = 0 \quad (18)$$

where B_i and A_i are transformation matrices from $\xi_i \eta_i \zeta_i$ to $x_i y_i z_i$ frames and from $x_i y_i z_i$ to XYZ frames, respectively. The same notations are used in defining matrices B_j and A_j for body j .

(c) Revolute Joint (RVLT). A revolute joint between adjacent bodies i and j is shown in Fig. 4. Three points, P in both bodies, I in body i , and J in body j , are chosen on the common joint axes. This joint requires that point P should be common to both bodies and that vectors g_i and g_j must be parallel. Hence, point P must satisfy the spherical joint equation of Eq. 16.

Constraint equations that require vectors g_i and g_j in Fig. 4 to be parallel are that g_j be orthogonal to two perpendicular vectors h_1 and h_2 that are imbedded in body i and are perpendicular to g_i . Those constraints can be formulated as two DOT1 constraints,

$$\left. \begin{aligned} h_1^T g_j &= 0 \\ h_2^T g_j &= 0 \end{aligned} \right\} \quad (19)$$

VIII. SELECTION OF DEFORMATION MODES. In modal deformation approximate methods, a few of the lowest frequency natural vibration modes are normally used. This often requires a large number of vibration modes (i.e., high frequency modes) to represent local deformation effects due to concentrated loads. Developments in structural dynamics [20-22] suggest that another kind of deformation mode should be introduced to approximate the effects of high frequency modes.

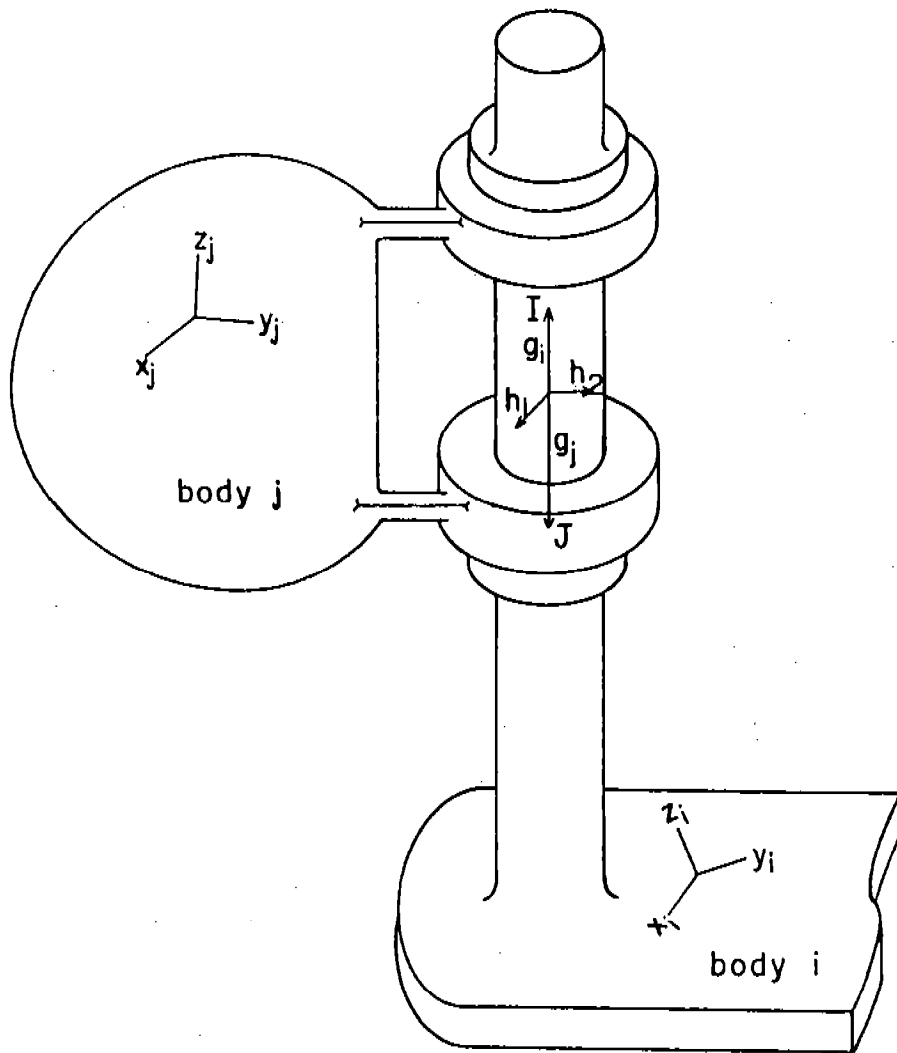


Figure 4. Revolute Joint

In mechanical system dynamics, there are differences from conventional structural dynamics. Kinematic constraints and nonlinear spring-damper elements are frequently attached between two adjacent elastic bodies and small elastic deformation is highly coupled with the geometrically nonlinear global motion of the system. To consider the local deformation due to forces transmitted by these elements, deformation modes are defined by imposing unit forces in the direction of reaction or spring forces that are expected to cause significant deformation of the body. The resulting deformation is called a static correction mode. A detailed derivation of static correction modes for application to machine dynamics can be found in Ref. 18.

If one chooses \bar{n} vibration normal modes from vibration analysis and \bar{m} static correction modes, the elastic displacement u can be written as

$$u = \Psi_{\bar{n}} a_{\bar{n}} + \Psi_{\bar{m}} a_{\bar{m}} \quad (20)$$

where $\Psi_{\bar{n}}$ and $\Psi_{\bar{m}}$ are mode shapes from vibration and static analysis, respectively, and $a_{\bar{n}}$ and $a_{\bar{m}}$ are amplitudes of these modes. The matrices $\Psi_{\bar{n}}$ and $\Psi_{\bar{m}}$ are combined to form

$$\Psi = \begin{pmatrix} \Psi_{\bar{n}} & \Psi_{\bar{m}} \end{pmatrix} \quad (21)$$

IX. COMPUTER IMPLEMENTATION. To practically implement the theory presented in this paper, a finite element code with the capability of carrying out both vibration and static analysis is employed. Most finite element codes can be used for this purpose. The SPAR Structural Analysis System was used for applications presented in this paper.

After vibration and static correction mode analysis is complete, the modal matrix is composed of vibration normal modes and static correction modes (Eq. 21). An intermediate processor is used to generate modal matrices and time-independent terms in the equations of motion. Output from the intermediate processor is read as input data in the DADS dynamic analysis code. Using the input data, the DADS program formulates the equations of motion of the system (Eq. 14), which are then integrated using a variable-step, variable-order numerical integration algorithm. For additional details regarding implementation, see Refs. 18 and 19.

X. NUMERICAL EXAMPLES.

(a) Flexible Component Door Closing Mechanism. A flexible door is attached to a rigid ground by two beams and revolute joints, as shown in Fig. 5. If the rotational axes of the revolute joints are parallel to the Z-axis, the door can freely rotate. If the revolute joints are not parallel, the door must deform to rotate. In this example, the rotational axes of the revolute joints remain in the X-Z plane and make an angle of 5 degrees with the Z-axis. The door is initially rotated 15 degrees. To make the body of the door much stiffer than the beams, the Modulus of Elasticity of the door plate is assigned to be 10 times greater than that of the beams. Material properties and dimensions of the mechanism are as follows;

Mass density, $\rho = 7.83 \times 10^3 \text{ Kg/m}^3$
 Height of the plate, $H = 0.8 \text{ m}$
 Width of the plate, $W = 0.6 \text{ m}$
 Modulus of Elasticity of the plate, $E_2 = 2.0 \times 10^{12} \text{ N/m}^2$
 Thickness of the plate, $t = 0.005 \text{ m}$
 Length of the beams, $L = 0.05 \text{ m}$
 Modulus of Elasticity of the beams, $E_1 = 2.0 \times 10^{11} \text{ N/m}^2$
 Radii of the beams, $r = 0.003 \text{ m}$.

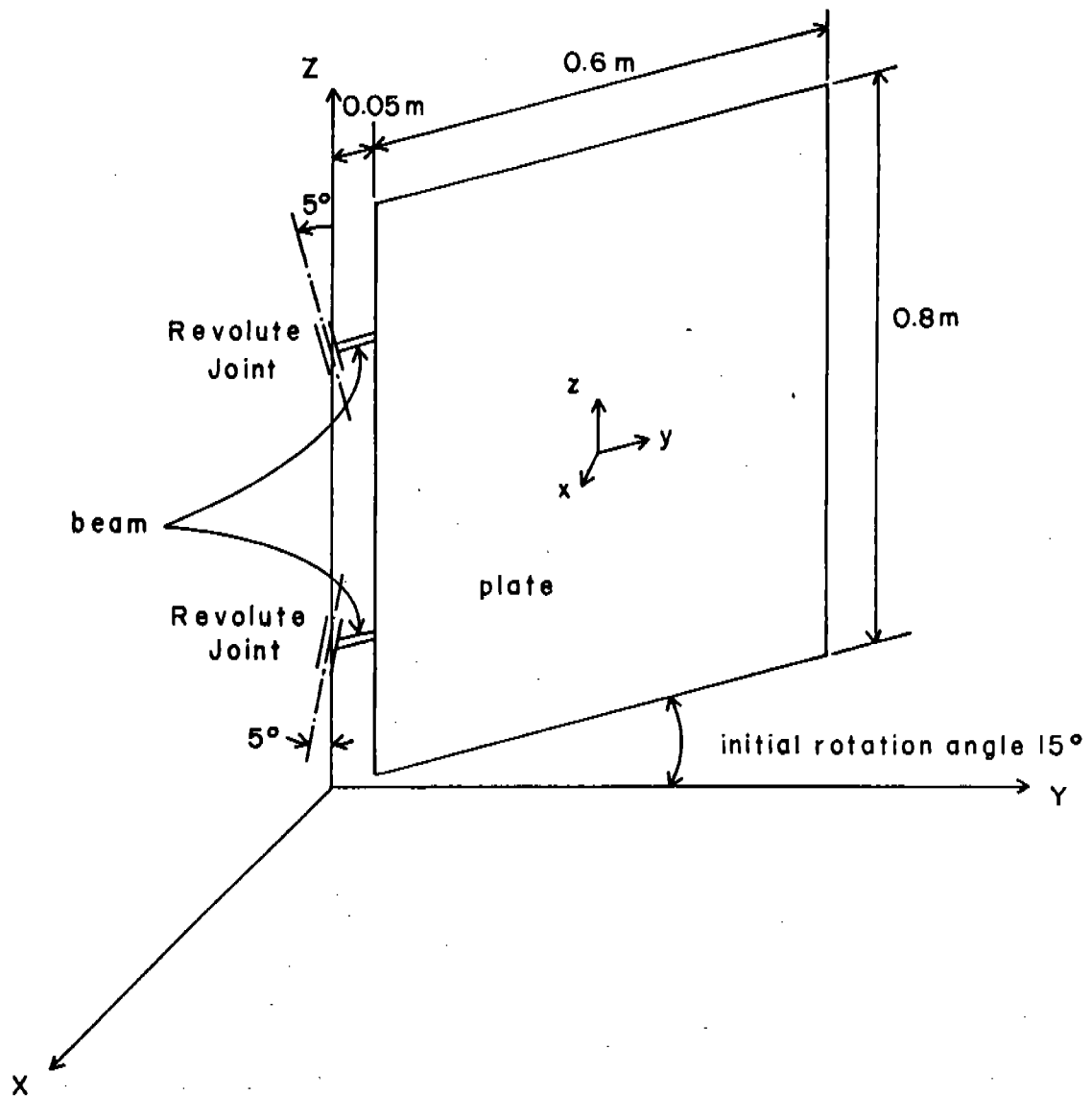
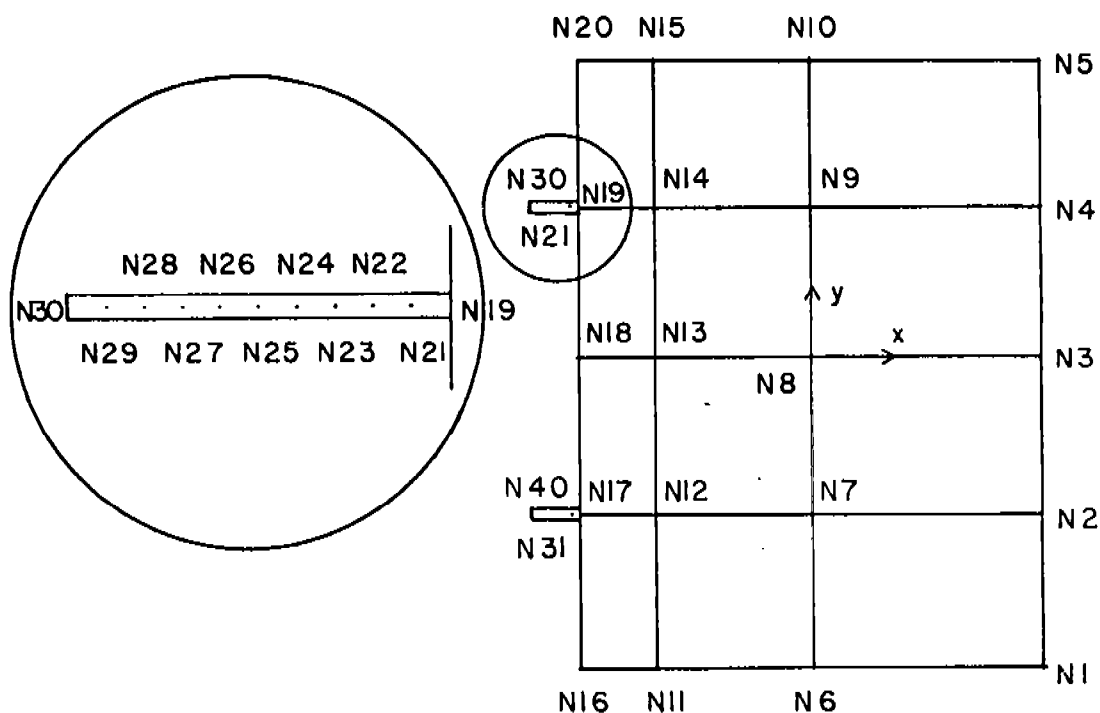


Figure 5. Flexible Door With Initial Rotation of Angle 15°

The finite element model of the door is shown in Fig. 6. The plate is divided into 12 membrane+bending elements and each beam is divided into 10 beam elements of equal length. Boundary conditions for vibration analysis are chosen such that three translational coordinates at node 40 are fixed and X- and Z-direction translational coordinates at node 30 are fixed. Since only five constraints are imposed for vibration analysis, there is one rigid body mode of vibration. The first ten natural frequencies from vibration analysis are 0.0, 62.08, 126.29, 163.53, 227.95, 237.84, 375.90, 477.91, 499.10, and 533.48 rad/sec.



N ** : Node Number **

Plate ; 12 (Membrane + Bending) Elements

$$E = 2.0 \times 10^{12} \text{ N/m}^2$$

Beam ; 10 Beam Elements at each beam

$$E = 2.0 \times 10^{11} \text{ N/m}^2$$

Figure 6. Finite Element Model of Flexible Door

Since the five constraints for vibration analysis are not sufficient to suppress rigid body motion, one additional constraint is imposed to define static correction modes. The Z-direction translational coordinate at node 8 is fixed as one additional constraint. Since a revolute joint allows one relative rotational degree of freedom about its joint axis, there are five reaction components at each revolute joint. In this model, however, five reaction components (2 forces at node 30 and 3 forces at node 40) are already accounted for in vibration analysis. The other five reaction components (two torques at nodes 30 and 40 and one force at node 30) are chosen to define static correction modes, as shown in Fig. 7. At node 30, the first static correction mode is due to a unit torque in the Z-direction, the second is due to a unit force along the rotational axis,

and the third is due to a unit torque along the axis in the X-Y plane perpendicular to the rotational axis. At node 40, the first mode is due to a unit torque in the Z-direction and the second is due to a unit torque along the axis in the X-Y plane perpendicular to the rotational axis.

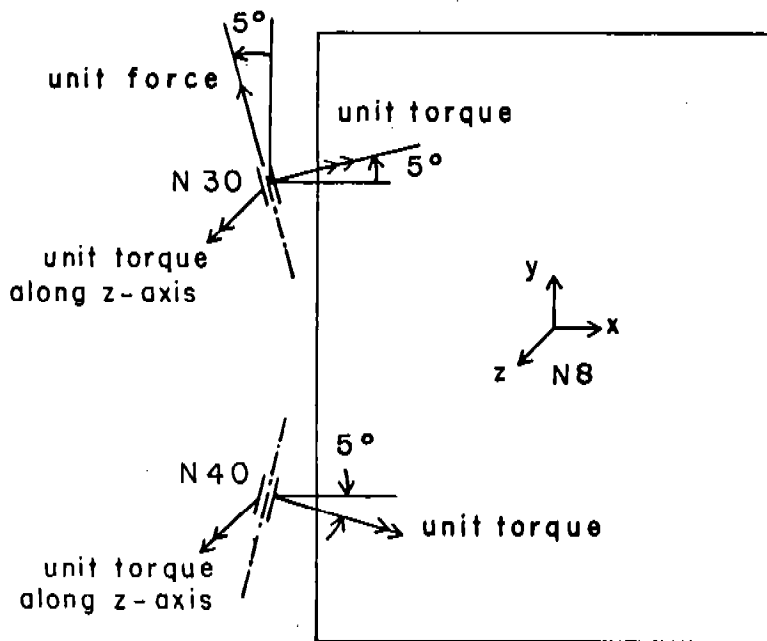


Figure 7. Forces Defining 5 Static Correction Modes

To see the vibratory motion of the door due to an initial 15 degree rotation, the door is released from the deformed position. Analysis was carried out with 5N (5-normal modes), 5S (5-static correction modes), 9N (9-normal modes), and 4N5S (4-normal + 5-static correction modes) models. The static correction modes in the 5S and 4N5S solutions are inertia relief and residual inertia relief attachment modes, respectively. Inertia relief attachment modes are defined by subtracting a portion of the rigid body mode from the attachment modes. Residual inertia relief attachment modes are then defined by subtracting the contribution of the kept normal modes from these inertia relief attachment modes.

The X-coordinates of node 8 in the 5N and 5S solutions are shown in Fig. 8. To determine the reason for the difference in frequency between the 5N and 5S solutions, strain energies at the initially deformed states are compared. The initial strain energy is calculated as $S.E. = (a^T K_{aa} a)/2$, where a is the vector of modal coordinates at the initially deformed configuration and K_{aa} is the modal stiffness matrix. The initial strain energy of the 5N state is 57.6 N·m and that of the 5S state is 0.562 N·m. Because of the high initial strain energy in the 5N state, the frequency of the 5N solution is much higher than that of the 5S solution. Considering the minimum potential energy principle, for the initial

equilibrium configuration, it can be seen that the 5S deformed state is closer to the actual initial deformation than that of 5N state. Thus, the 5S solution is expected to be more accurate than the 5N solution.

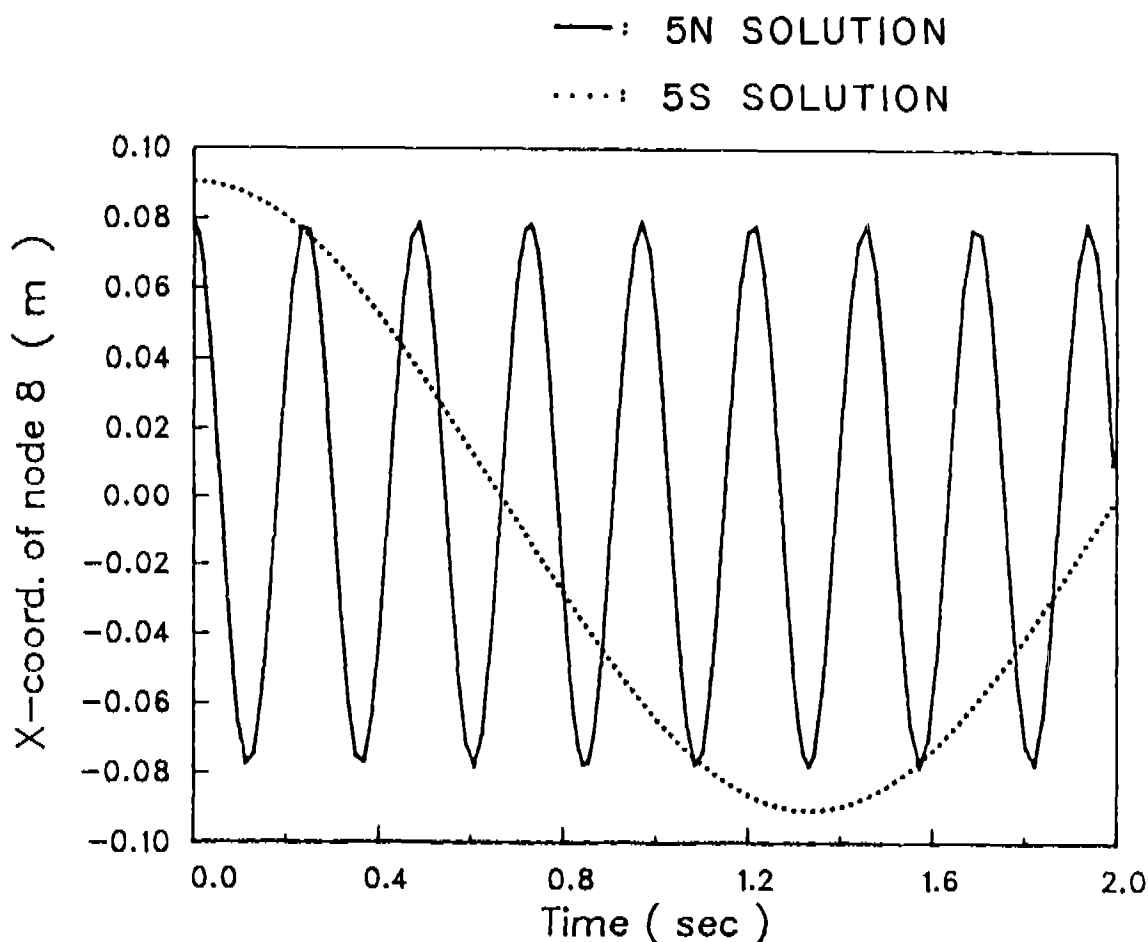


Figure 8. X-coordinates of Mode 8 in the 5N and 5S Solutions

Nine normal mode (9N) and 4N5S solutions are compared with the 5S solution in Fig. 9. Although the number of normal modes is increased to nine in the 9N solution, the solution is still far from the 5S solution. This means that nine normal modes are not sufficient in this example. There is almost no difference between results of the 5S and 4N5S solutions. This means that the attachment modes are dominant in this example.

Simulation times on a PRIME 750 computer and RMS (Root Mean Square) integration stepsizes are given in Table 1. Since the frequency of the normal mode solution is much greater than that of the 5S solution, its computing time is extreme. It is difficult to estimate the number of normal modes that would be required for an accurate solution. Certainly, computing time would be astronomical.

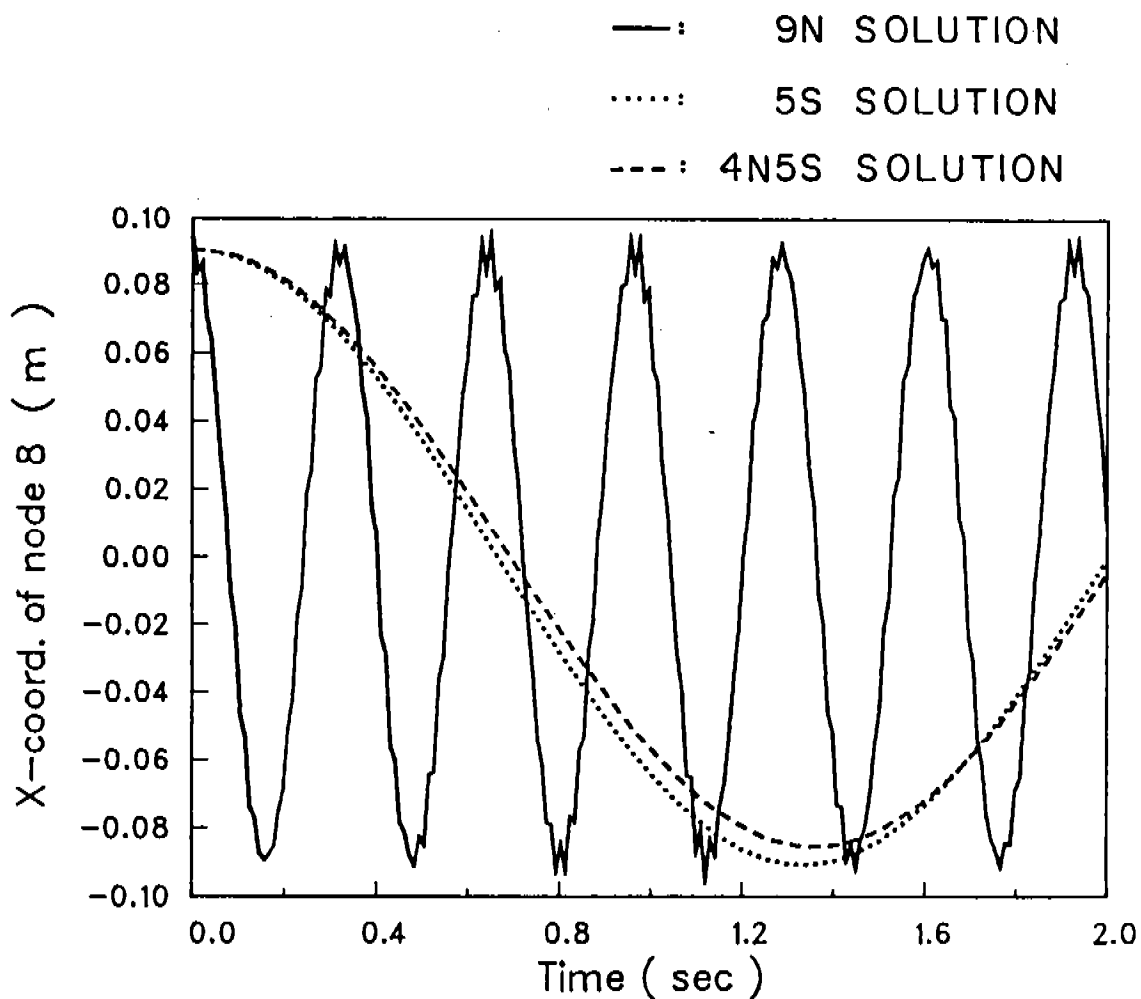


Figure 9. X-coordinates of Node 8 in the 9N, 5S, and 4N5S Solutions

Table 1. Comparison of Simulation Times

Type	T end [sec]	CPU [sec]	RMS integration stepsize [sec]
5S solution	2.0	106	0.47812E-01
5N solution	2.0	401	0.90934E-02
9N solution	2.0	7471	0.65035E-03
4N5S solution	2.0	7281	0.75491E-03

(b) Windshield Wiper Mechanism. A model of a two blade windshield wiper mechanism that consists of 6 bodies is shown in Fig. 10. The crank arm (body 2), drive link (body 3), and connecting link (body 5) are modeled as rigid bodies. In the right wiper arm (body 4), the link connecting bar and wiper arm are modeled as rigid and flexible, respectively. In the left wiper arm (body 6), the link connecting bar and wiper are modeled as rigid and flexible, respectively. The chassis of the vehicle is body 1. The mass and moments of inertia of each body are given in Table 2.

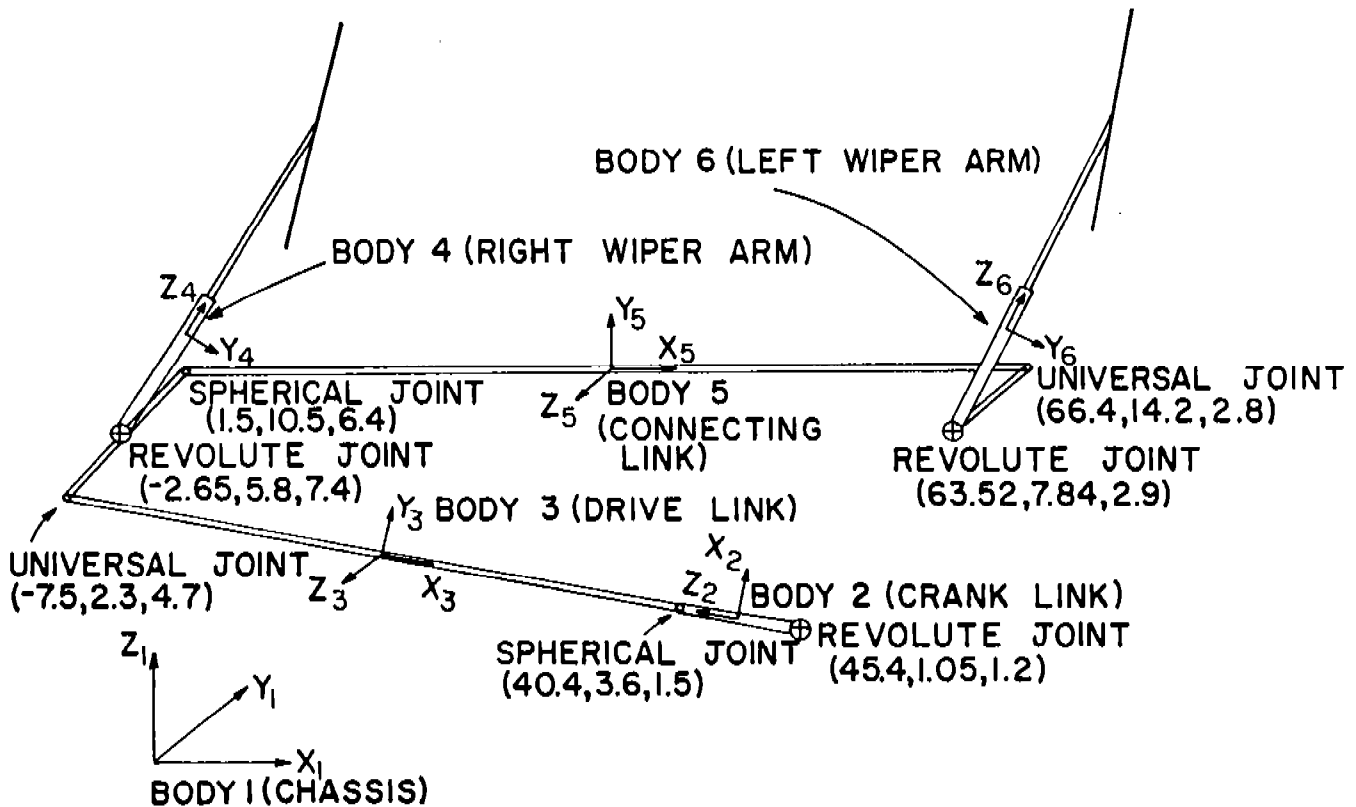


Figure 10. Windshield Wiper Mechanism

Table 2. Mass and Moments of Inertia

Body No.	Mass (gram)	Moments of Inertia ($\text{g} \cdot \text{cm}^2$)					
		I_{xx}	I_{yy}	I_{zz}	I_{xy}	I_{xz}	I_{yz}
2	61.3	282.6	314.7	33.6	0.0	0.0	0.0
3	189.2	100.0	36413.8	36413.8	0.0	0.0	0.0
4	567.0	215730.0	218240.0	3610.6	155.5	-1737.8	-4060.0
5	255.9	300.0	90133.7	90133.7	0.0	0.0	0.0
6	499.5	198920.0	202210.0	4173.8	-486.6	10285.0	-4006.0

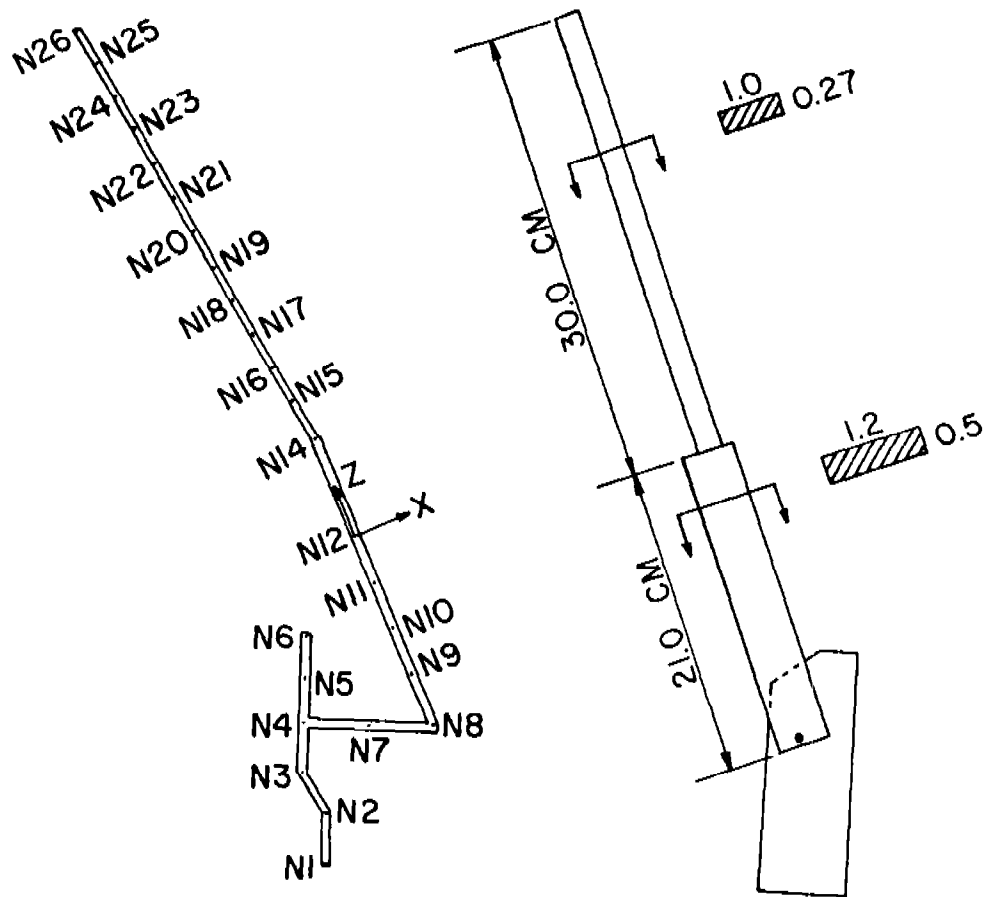
There are three revolute joints, two spherical joints, and two universal joints in the model. Locations of joints at extreme position measured in the chassis coordinate system are shown in Fig. 10. The first revolute joint is located between the chassis and crank arm. The second and third revolute joints are located between the chassis and the right and left link connecting bars, respectively. The first spherical joint is located between the crank arm and drive link. The second spherical joint is between the right link connecting bar and connecting link. The first universal joint is located between the drive link and the right link connecting bar. The second universal joint is between the connecting link and left link connecting bar.

A finite element model of the right and left wiper arms is shown in Fig. 11. This model consists of 26 nodes. The link connecting bar (nodes 1-8) is treated as rigid and the wiper arm (nodes 8-26) is treated as flexible. The cross-section of the wiper arm is modeled as rectangular, with two different shapes along its length. To consider the mass of the wiper blade in the finite element model, a nonstructural mass element is attached at node 26. Nonstructural mass elements are also allocated between nodes 8 and 14, to account for variation of the cross-sections. To calculate the center of mass of the arm, the mass of the link connecting bar is lumped at its nodes. To consider elastic deformation of the wiper arm, due to frictional force at the blade, one static correction mode is defined and used in flexible body analysis. That static correction mode is defined by a unit force applied at the tip, perpendicular to the wiper arm.

The friction force on the arm tip is modeled as a linearly decreasing force with increasing tip velocity, as shown in Fig. 12. Simulations are carried out with a constant crank arm speed of 60 rpm. Tip velocity and friction force from rigid and flexible solutions are presented as plots of variables versus time in Fig. 13. As shown in Fig. 13, the linear tip velocities in both cases are different. In the rigid case, vibratory phenomenon can not be detected, but a vibratory frequency of 13-14 Hz is predicted in the flexible case.

XI. CONCLUSIONS. This paper presents a new method of choosing modal matrices for simulation of flexible mechanical systems, employing static correction and vibration modes. The following conclusions are obtained.

- (1) When high reaction forces occur at kinematic joints, static correction modes are much better than vibration modes to represent deformation.
- (2) If vibration and static correction modes are combined, deformation due to both inertial and reaction forces can be accurately represented.
- (3) Computing cost can be significantly reduced if static correction modes are employed, instead of purely normal modes of vibration.



N** : NODE NUMBER
BEAM ELEMENTS

Figure 11. Finite Element Model of Wiper Arms

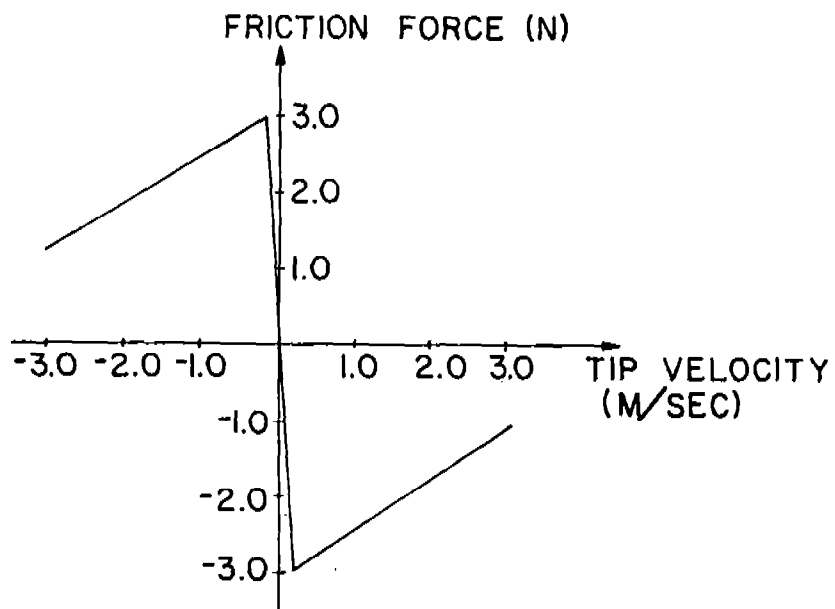


Figure 12. Friction Force Versus Tip Velocity

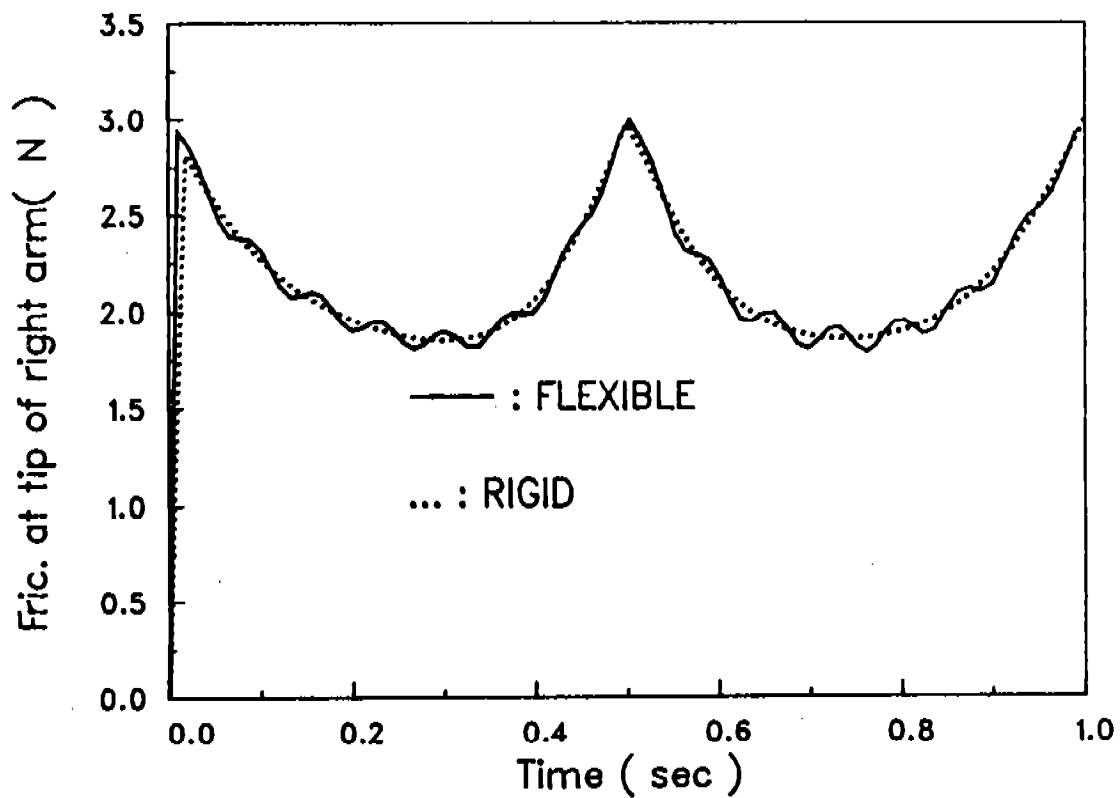
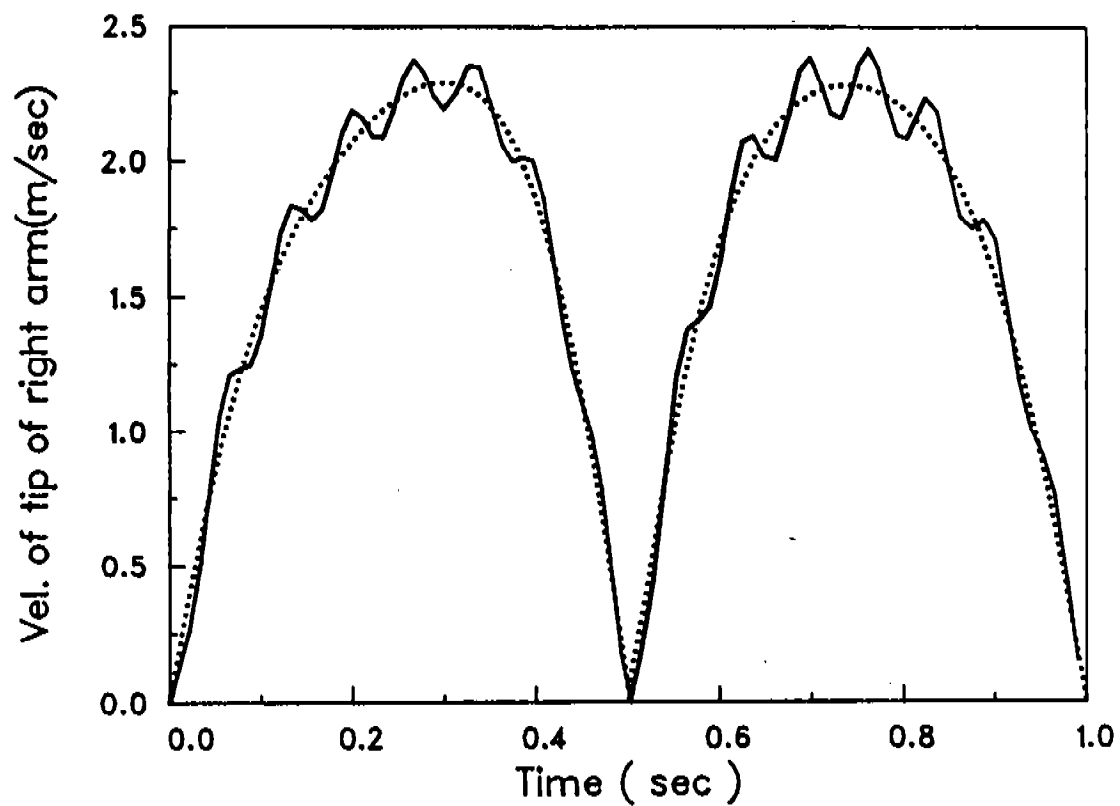


Figure 13. Velocity and Friction Force at Tip of Right Arm

ACKNOWLEDGMENT

This research was sponsored by U.S. Army Research Office Grant No. DAAG 29-82K-0086 and by the Delco Products Division of General Motors Corporation.

REFERENCES

1. Winfrey, R.C., "Elastic Link Mechanism Dynamics," Journal of Engineering for Industry, Vol. 93, No. 1, February 1971, pp. 268-272.
2. Winfrey, R.C., "Dynamic Analysis of Elastic Link Mechanisms by Reduction of Coordinates," Journal of Engineering for Industry, Vol. 94, No. 2, May 1972, pp. 577-582.
3. Erdman, A.G., and Sandor, G.N., Advanced Mechanism Design; Analysis and Synthesis, Vol. II, Prentice-Hall, Englewood Cliffs, N.J., 1984.
4. Sadler, J.P., and Sandor, G.N., "A Lumped Parameter Approach to Vibration and Stress Analysis of Elastic Linkages," Journal of Engineering for Industry, Vol. 95, No. 2, May 1973, pp. 549-557.
5. Bahgat, B.M., and Willmert, K.D., "Finite Element Vibrational Analysis of Planar Mechanisms," Mechanism and Machine Theory, Vol. 11, 1976, pp. 47-71.
6. Chu, S.C., and Pan, K.C., "Dynamic Response of a High-Speed Slider-Crank Mechanism with an Elastic Connecting Rod," Journal of Engineering for Industry, Vol. 97, No. 2, May 1975, pp. 542-550.
7. Likins, P., "Dynamic Analysis of a System of Hinge-Connected Rigid Bodies with Nonrigid Appendage," International Journal of Solids and Structures, Vol. 9, 1973, pp. 1473-1487.
8. Hooker, W., "Equations of Motion for Interconnected Rigid and Elastic Bodies: A Derivation Independent of Angular Momentum," Celestial Mechanics, 11, 1975, pp. 337-359.
9. Ho, J., "Direct Path Method for Flexible Multibody Spacecraft Dynamics," Journal of Spacecraft and Rockets, Vol. 14, Feb. 1977, pp. 102-110.
10. Bodley, C., Devers, A., Park, A., and Frisch, H., A Digital Computer Program for Dynamic Interaction Simulation of Controls and Structures (DISCOS), NASA Technical Paper 1219, May 1978.
11. Boland, P., Samin, J.C., and Willems, P.Y., "Stabiity Analysis of Interconnected Deformable Bodies with Closed-Loop Configuration", AIAA Journal, Vol. 13, 1975, pp. 864-867

12. Jerkovsky, W., "The Structure of Multibody Dynamics Equations", Journal of Guidance and Control, Vol. 1, No.3, 1978, pp. 173-182
13. Keat, J., Dynamical Equations of Multi-Body Systems with Application to Space Structures Deployment, Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1983
14. Song, J.O., and Haug, E.J., "Dynamic Analysis of Planar Flexible Mechanisms," Computer Methods in Applied Mechanics and Engineering, Vol. 24, 1980, pp. 359-381.
15. Sunada, W., and Dubowsky, S., "The Application of Finite Element Methods to the Dynamic Analysis of Flexible Spatial and Co-Planar Linkage Systems," Journal of Mechanical Design, Vol. 103, July 1981, pp. 643-651.
16. Shabana, A.A., and Wehage, R.A., "A Coordinate Reduction Technique for Transient Analysis of Spatial Structures with large Angular Rotations," Journal of Structural Mechanics, Vol. 11, No. 3, 1983, pp. 401-431.
17. Wittenburg, I.J., "Dynamics of Systems of Rigid Bodies," B.G. Teuber, Stuttgart, 1977.
18. Yoo, W.S., and Haug, E.J., "Dynamics of Articulated Structures, Part I: Theory", Journal of Structural Mechanics, to appear.
19. Yoo, W.S., and Haug, E.J., "Dynamics of Articulated Structures, Part II: Computer Implementation and Applications", Journal of Structural Mechanics, to appear.
20. MacNeal, R.H., "A Hybrid Method of Component Mode Synthesis", Computer & Structures, Vol. 1, 1971, pp. 581-601.
21. Craig, R.R. and Chang, C.J., "On the Use of Attachment Modes in Substructure Coupling for Dynamic Analysis", Dynamics & Structural Dynamics, AIAA/ASME 18th Structures, Structural Dynamics & Material Conference, 1977.
22. Craig, R.R., "Component Mode Synthesis", Structural Dynamics : An Introduction to Computer Methods, Wiley 1981, pp. 467-495.
23. Wehage, R.A. and Haug, E.J., "Generalized Coordinate Partitioning for Dimension Reduction in Analysis of Constrained Dynamic Systems", Journal of Mechanical Design, Vol. 104, Jan. 1982, pp. 247-255.

APPENDIX

The following matrix notation is used to implement vector operations:

$$a \equiv [a_x, a_y, a_z]^T$$

$$\tilde{a} = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix}$$

$$a \cdot b \equiv a^T b, \quad a \times b \equiv \tilde{a} b$$

$$^+a \equiv \begin{bmatrix} 0 & a^T \\ -a & \tilde{a} \end{bmatrix}$$

Using Euler's theorem, Euler parameters [17] can be defined as

$$e_0 = \cos \frac{\chi}{2}, \quad e = n \sin \frac{\chi}{2}$$

where χ is the rotation angle about a unit vector n that moves the body

from a reference orientation to a general orientation and

$e \equiv [e_1, e_2, e_3]^T$. Defining $p \equiv [e_0, e_1, e_2, e_3]^T$, and has the following relations:

$$p^T p = 1, \quad p^T \dot{p} = \dot{p}^T p = 0$$

$$E \equiv [-e, \tilde{e} + e_0 I], \quad G \equiv [-e, -\tilde{e} + e_0 I]$$

$$E^T E = G^T G = -pp^T + I_{4 \times 4}, \quad G \dot{a} = \tilde{a} G - a p^T$$

$$E p = G p = 0, \quad E E^T = G G^T = I, \quad E G^T = A$$

RESPONSE OF DAMPED MECHANICAL SYSTEMS
TO A TIME DEPENDENT EXTERNAL FORCE

Gary L. Anderson
Engineering Sciences Division
US Army Research Office
Research Triangle Park, North Carolina

ABSTRACT. The method of Krylov and Bogoliubov is extended so as to be applicable to ordinary differential equations that describe the motion of linearly damped, linear and nonlinear mechanical systems that are subjected to a particular family of oscillatory external forces. These force functions, in turn, satisfy a linear second order ordinary differential equation with constant coefficients. This method is particularly useful for the resonance analysis of systems in transient or steady motion. The technique is applied to nonhomogeneous forms of the differential equations of Mathieu, Bessel, Hermite and Duffing for the purpose of generating approximate analytic expressions for their solution. Numerical computations based upon these analytical approximations and the Runge-Kutta numerical integration technique reveal that the method developed here produces rather useful accurate approximations.

I. INTRODUCTION. The classical method of averaging as developed by Krylov, Bogoliubov and Mitropolsky [1], [2] has been applied by many investigators to solve linear and nonlinear differential equations of the type

$$\ddot{x} + \omega^2 x + \epsilon f(t, x, \dot{x}) = 0, \quad (1)$$

which arise in the theories of mechanical and electrical oscillations. In order to account for the influence of linear damping, several authors have considered the differential equation

$$\ddot{x} + 2\gamma\dot{x} + \omega^2x + \epsilon f(t, x, \dot{x}) = 0, \quad (2)$$

and have introduced certain modifications of the Krylov-Bogoliubov theory. Various such techniques have been reported by Brunelle [3], Mendelson [4] and Anderson [5] to [8]. Stanasic and Euler [9] have extended the Krylov-Bogoliubov method to the specific non-linear forced motion problem

$$\ddot{x} + \omega^2x + \mu_1x^3 + \mu_2x\dot{x}^2 = N\cos\Omega t, \quad (3)$$

which is an equation that describes the motion of a large class of nonlinear mechanical systems, as reported by Kane [10]. Stanasic and Euler have not accounted for the possible presence of a linear damping term in (3).

The Stanasic-Euler technique consists of reducing the nonhomogeneous, nonlinear equation in (3) to a fourth order homogeneous, non-linear equation. In the form of the solution assumed, there appear two variable amplitudes and two variable frequencies that are to be determined.

In the present investigation, the basic approach of the Stanasic-Euler solution is extended to a more general class of problems, namely,

$$\ddot{x} + 2\gamma_1\dot{x} + \omega_1^2x + \epsilon f(t, x, \dot{x}) = \epsilon Q(t), \quad (4)$$

where $Q(t)$ is known to satisfy the linear differential equation

$$\ddot{Q}(t) + 2\gamma_2\dot{Q}(t) + \omega_2^2Q(t) = 0. \quad (5)$$

The initial conditions associated with (4) are taken to be

$$x(t_0) = x_0, \quad \dot{x}(t_0) = v_0 \quad (6)$$

Furthermore, it is assumed that $\gamma_j < \omega_j$, $j = 1, 2$, so that the system is underdamped in the linear approximation. The problem of determining the response of

dynamic systems described by (4) to (6) is of considerable importance for the design of nonlinear elements in machines, mechanisms, vehicles and other structures.

2. METHOD OF SOLUTION. The objective of the analysis that follows is to determine an approximate solution for the initial value problem stated in (4) and (6). The forcing function $Q(t)$ appearing in (4) is assumed to satisfy the linear differential equation in (5). The first step in the derivation of an approximation for $x(t)$ is the elimination of the presence of $Q(t)$ in the equation of motion. This is done by forming the first and second derivatives with respect to time t of (4) and substituting the results into (5). This process leads to the following fourth order homogeneous nonlinear differential equation:

$$x^{iv} + 2(\gamma_1 + \gamma_2)\ddot{x} + (\omega_1^2 + 4\gamma_1\gamma_2 + \omega_2^2)\ddot{x} + 2(\gamma_2\omega_1^2 + \gamma_1\omega_2^2)\dot{x} + (\omega_1\omega_2)^2x + \epsilon F(t, x, \dot{x}, \ddot{x}, \ddot{\ddot{x}}) = 0, \quad (7)$$

where

$$\begin{aligned} F(t, x, \dot{x}, \ddot{x}, \ddot{\ddot{x}}) = & \frac{\partial^2 f}{\partial t^2} + 2\gamma_2 \frac{\partial f}{\partial t} + \omega_2^2 f + 2\left[\frac{\partial^2 f}{\partial x \partial t} + \gamma_2 \frac{\partial f}{\partial x}\right]\dot{x} + \\ & \left[\frac{\partial f}{\partial x} + 2\frac{\partial^2 f}{\partial x \partial t} + 2\gamma_2 \frac{\partial f}{\partial x}\right]\ddot{x} + \frac{\partial f}{\partial \ddot{x}}\ddot{\ddot{x}} + \frac{\partial^2 f}{\partial x^2}\dot{x}^2 + \\ & + 2\frac{\partial^2 f}{\partial x \partial \ddot{x}}\ddot{x}\ddot{x} + \frac{\partial^2 f}{\partial \ddot{x}^2}\ddot{\ddot{x}}^2. \end{aligned} \quad (8)$$

Two initial conditions for (7) are already available in (6), but two more must be adjoined to these. From (4) and its derivative with respect to t , we find that these are

$$\ddot{x}(t_0) = p_0, \quad \ddot{x}(t_0) = q_0 \quad (9)$$

where

$$p_0 = \epsilon Q(t_0) - \omega_1^2 x_0 - 2\gamma_1 v_0 - \epsilon f(t_0, x_0, v_0), \quad (10)$$

$$q_0 = \epsilon \dot{Q}(t_0) - \omega_1^2 v_0 - 2\gamma_1 p_0 - \epsilon \left[\frac{\partial f(t_0, x_0, v_0)}{\partial t} + v_0 \frac{\partial f(t_0, x_0, v_0)}{\partial x} + p_0 \frac{\partial f(t_0, x_0, v_0)}{\partial \dot{x}} \right]. \quad (11)$$

Motivated by the form of solution of the linear counterpart ($\epsilon = 0$) of (7), we assume that the solution of the original differential equation (4) can be expressed in the form

$$x(t) = a_1 \sin \psi_1 + a_2 \sin \psi_2, \quad (12)$$

where $a_j = a_j(t)$ and $\psi_j = \Omega_j t + \theta_j(t)$, $j = 1, 2$, $a_j(t)$ and $\theta_j(t)$ being unknown slowly varying amplitude and phase functions with

$$\Omega_j = (\omega_j^2 - \gamma_j^2)^{1/2}, \quad \gamma_j < \omega_j, \quad j = 1, 2. \quad (13)$$

If we form the first, second and third derivatives of (12) with respect to time and proceed in the spirit of the method of variations of parameters, we obtain

$$\dot{x} = \Omega_1 a_1 \cos \psi_1 - \gamma_1 a_1 \sin \psi_1 + \Omega_2 a_2 \cos \psi_2 - \gamma_2 a_2 \sin \psi_2, \quad (14)$$

$$\begin{aligned} \ddot{x} = & -2\gamma_1 \Omega_1 a_1 \cos \psi_1 - (\Omega_1^2 - \gamma_1^2) a_1 \sin \psi_1 - 2\gamma_2 \Omega_2 a_2 \cos \psi_2 - \\ & - (\Omega_2^2 - \gamma_2^2) a_2 \sin \psi_2, \end{aligned} \quad (15)$$

$$\ddot{x} = (3\gamma_1^2\Omega_1 - \Omega_1^3)a_1\cos\psi_1 + (3\gamma_1\Omega_1^2 - \gamma_1^3)a_1\sin\psi_1 + (3\gamma_2^2\Omega_2 - \Omega_2^3)a_2\cos\psi_2 + (3\gamma_2\Omega_2^2 - \gamma_2^3)a_2\sin\psi_2 \quad (16)$$

and

$$(\dot{a}_1 + \gamma_1 a_1)\sin\psi_1 + a_1\dot{\theta}_1\cos\psi_1 + (\dot{a}_2 + \gamma_2 a_2)\sin\psi_2 + a_2\dot{\theta}_2\cos\psi_2 = 0, \quad (17)$$

$$(\dot{a}_1 + \gamma_1 a_1)(\Omega_1\cos\psi_1 - \gamma_1\sin\psi_1) - a_1\dot{\theta}_1(\gamma_1\cos\psi_1 + \Omega_1\sin\psi_1) + (\dot{a}_2 + \gamma_2 a_2)(\Omega_2\cos\psi_2 - \gamma_2\sin\psi_2) - a_2\dot{\theta}_2(\gamma_2\cos\psi_2 + \Omega_2\sin\psi_2) = 0, \quad (18)$$

$$(\dot{a}_1 + \gamma_1 a_1)[2\gamma_1\Omega_1\cos\psi_1 + (\Omega_1^2 - \gamma_1^2)\sin\psi_1] - a_1\dot{\theta}_1[2\gamma_1\Omega_1\sin\psi_1 - (\Omega_1^2 - \gamma_1^2)\cos\psi_1] + (\dot{a}_2 + \gamma_2 a_2)[2\gamma_2\Omega_2\cos\psi_2 + (\Omega_2^2 - \gamma_2^2)\sin\psi_2] - a_2\dot{\theta}_2[2\gamma_2\Omega_2\sin\psi_2 - (\Omega_2^2 - \gamma_2^2)\cos\psi_2] = 0. \quad (19)$$

The specific forms retained in (14) to (16) are analogous to those that would occur in the analysis of the corresponding linear problem.

Substituting (14) to (16) and the apposite expression for x^{IV} into (7), we obtain

$$(\dot{a}_1 + \gamma_1 a_1)(e_1\cos\psi_1 + f_1\sin\psi_1) + a_1\dot{\theta}_1(f_1\cos\psi_1 - e_1\sin\psi_1) + (\dot{a}_2 + \gamma_2 a_2)(e_2\cos\psi_2 + f_2\sin\psi_2) + a_2\dot{\theta}_2(f_2\cos\psi_2 - e_2\sin\psi_2) = -\epsilon F, \quad (20)$$

where F has been defined in (8) and for $j = 1, 2$

$$e_j = \Omega_j(3\gamma_j^2 - \Omega_j^2), \quad f_j = \gamma_j(3\Omega_j^2 - \gamma_j^2)$$

The system of equations in (17) to (20) may be considered a system of linear algebraic equations in the four unknown $\dot{a}_j + \gamma_j a_j$ and $\dot{\theta}_j$, $j = 1, 2$. The process of solving this system of equations leads to the following system of first order differential equations:

$$\dot{a}_1 + \gamma_1 a_1 = \epsilon F(R_{11} \cos \psi_1 + R_{12} \sin \psi_1), \quad (21)$$

$$\dot{\theta}_1 = -(\epsilon F/a_1)(R_{11} \sin \psi_1 - R_{12} \cos \psi_1), \quad (22)$$

$$\dot{a}_2 + \gamma_2 a_2 = \epsilon F(R_{22} \cos \psi_2 - R_{12} \sin \psi_2), \quad (23)$$

$$\dot{\theta}_2 = -(\epsilon F/a_2)(R_{22} \sin \psi_2 + R_{12} \cos \psi_2), \quad (24)$$

where

$$R_{11} = [\omega_1^2 - \omega_2^2 - 2\gamma_1(\gamma_1 - \gamma_2)]/\Omega_1 R, \quad R_{12} = 2(\gamma_2 - \gamma_1)/R,$$

$$R_{22} = [\omega_2^2 - \omega_1^2 + 2\gamma_2(\gamma_1 - \gamma_2)]/\Omega_2 R, \quad (25)$$

$$R = (\omega_2^2 - \omega_1^2)^2 + 4(\gamma_1 \omega_2 + \gamma_2 \omega_1)^2 - 4\gamma_1 \gamma_2 (\omega_1 + \omega_2)^2.$$

The system of first order differential equations in (21) to (24) must be accompanied by a set of four initial conditions. According to (6), (9), (12) and (14) to (16), we have

$$a_{10} \sin \psi_{10} + a_{20} \sin \psi_{20} = x_0$$

$$\Omega_1 a_{10} \cos \psi_{10} - \gamma_1 a_{10} \sin \psi_{10} + \Omega_2 a_{20} \cos \psi_{20} - \gamma_2 a_{20} \sin \psi_{20} = v_0$$

$$2\gamma_1 \Omega_1 a_{10} \cos \psi_{10} + (\Omega_1^2 - \gamma_1^2) a_{10} \sin \psi_{10} + 2\gamma_2 \Omega_2 a_{20} \cos \psi_{20} + \quad (26)$$

$$+ (\Omega_2^2 - \gamma_2^2) a_{20} \sin \psi_{20} = -p_0,$$

$$\Omega_1 (3\gamma_1^2 - \Omega_1^2) a_{10} \cos \psi_{10} + \gamma_1 (3\Omega_1^2 - \gamma_1^2) a_{10} \sin \psi_{10} + \Omega_2 (3\gamma_2^2 - \Omega_2^2) a_{20} \cos \psi_{20} +$$

$$+ \gamma_2 (3\Omega_2^2 - \gamma_2^2) a_{20} \sin \psi_{20} = q_0,$$

where $a_{j0} = a_j(t_0)$ and $\psi_{j0} = \omega_j t_0 + \theta_j(t_0)$, $j = 1, 2$. This is a system of four algebraic equations in the four unknowns $a_{j0} \cos \psi_{j0}$ and $a_{j0} \sin \psi_{j0}$. Solving this system for $a_j(t_0)$, $\theta_j(t_0)$, we find

$$\begin{aligned} a_1(t_0) &= (W_1^2/\Omega_1^2 + W_2^2)^{1/2}/R, \\ \theta_1(t_0) &= \tan^{-1}(\Omega_1 W_2/W_1) - \omega_1 t_0, \\ a_2(t_0) &= -(Z_1^2/\Omega_2^2 + Z_2^2)^{1/2}/R, \\ \theta_2(t_0) &= \tan^{-1}(\Omega_2 Z_2/Z_1) - \omega_2 t_0, \end{aligned} \quad (27)$$

where

$$\begin{aligned} W_1 &= x_0 \omega_2^2 [4\gamma_1^2(\gamma_1 - \gamma_2) + (2\gamma_2 - 3\gamma_1)\omega_1^2 + \gamma_1 \omega_2^2] + v_0 [\omega_2^2(\omega_2^2 - \omega_1^2) + \\ &\quad + 2\gamma_2(2\gamma_2 - 3\gamma_1)\omega_1^2 + 2\gamma_1^2 \omega_2^2 + 8\gamma_1^2 \gamma_2(\gamma_1 - \gamma_2)] + p_0 [(\gamma_1 + 2\gamma_2)\omega_2^2 - \\ &\quad - 3\gamma_1 \omega_1^2 + 4\gamma_1(\gamma_1^2 - \gamma_2^2)] + q_0 [\omega_2^2 - \omega_1^2 + 2\gamma_1(\gamma_1 - \gamma_2)], \\ W_2 &= x_0 \omega_2^2 [\omega_2^2 - \omega_1^2 + 4\gamma_1(\gamma_1 - \gamma_2)] + 2v_0 [\gamma_1 \omega_2^2 - \gamma_2 \omega_1^2 + 4\gamma_1 \gamma_2(\gamma_1 - \gamma_2)] + \\ &\quad + p_0 [\omega_2^2 - \omega_1^2 + 4(\gamma_1^2 - \gamma_2^2)] + 2q_0(\gamma_1 - \gamma_2), \quad (28)_1 \\ Z_1 &= x_0 \omega_1^2 [(2\gamma_1 - 3\gamma_2)\omega_2^2 + \gamma_2 \omega_1^2 + 4\gamma_2^2(\gamma_2 - \gamma_1)] + v_0 [\omega_1^2(\omega_1^2 - \omega_2^2) + \\ &\quad + 2\gamma_2^2 \omega_1^2 + 2\gamma_1(2\gamma_1 - 3\gamma_2)\omega_2^2 + 8\gamma_1 \gamma_2^2(\gamma_2 - \gamma_1)] + p_0 [(\gamma_2 + 2\gamma_1)\omega_1^2 - \\ &\quad - 3\gamma_2 \omega_2^2 + 4\gamma_2(\gamma_2^2 - \gamma_1^2)] + q_0 [\omega_1^2 - \omega_2^2 + 2\gamma_2(\gamma_2 - \gamma_1)], \\ Z_2 &= x_0 \omega_1^2 [\omega_1^2 - \omega_2^2 + 4\gamma_2(\gamma_2 - \gamma_1)] + 2v_0 [\gamma_2 \omega_1^2 - \gamma_1 \omega_2^2 + 4\gamma_1 \gamma_2(\gamma_2 - \gamma_1)] + \\ &\quad + p_0 [\omega_1^2 - \omega_2^2 + 4(\gamma_2^2 - \gamma_1^2)] + 2q_0(\gamma_2 - \gamma_1). \end{aligned}$$

In the event that $\gamma_1 = \gamma_2 = 0$, i.e., there is no damping in the system, the forms of the initial conditions in (27) become significantly simpler. Using (6), (10), (11) and (27), we can verify that

$$\begin{aligned} a_1(t_0) &= [(v_0\omega_2^2 + q_0)^2/\omega_1^2 + (x_0\omega_2^2 + p_0)^2]^{1/2}/|\omega_2^2 - \omega_1^2|, \\ \theta_1(t_0) &= \tan^{-1} \left[\frac{\omega_1(x_0\omega_2^2 + p_0)}{v_0\omega_2^2 + q_0} \right] - \omega_1 t_0, \\ a_2(t_0) &= [(v_0\omega_1^2 + q_0)^2/\omega_2^2 + (x_0\omega_1^2 + p_0)^2]^{1/2}/|\omega_2^2 - \omega_1^2|, \\ \theta_2(t_0) &= \tan^{-1} \left[\frac{\omega_2(x_0\omega_1^2 + p_0)}{v_0\omega_1^2 + q_0} \right] - \omega_2 t_0. \end{aligned} \quad (28)_2$$

In the spirit of the Krylov-Bogoliubov method [1], we form the averages of (21) to (24) to obtain

$$\begin{aligned} \dot{a}_1 + \gamma_1 a_1 &= \varepsilon [R_{11} \langle F \cos \psi_1 \rangle + R_{12} \langle F \sin \psi_1 \rangle], \\ \dot{\theta}_1 &= -(\varepsilon/a_1) [R_{11} \langle F \sin \psi_1 \rangle - R_{12} \langle F \cos \psi_1 \rangle], \\ \dot{a}_2 + \gamma_2 a_2 &= \varepsilon [R_{22} \langle F \cos \psi_2 \rangle - R_{12} \langle F \sin \psi_2 \rangle], \\ \dot{\theta}_2 &= (\varepsilon/a_2) [R_{22} \langle F \sin \psi_2 \rangle + R_{12} \langle F \cos \psi_2 \rangle], \end{aligned} \quad (29)$$

$$\begin{aligned} \text{where } \langle F \cos \psi_j \rangle &= (1/4\pi^2) \int_0^{2\pi} \int_0^{2\pi} F \cos \psi_j d\psi_1 d\psi_2, \\ \langle F \sin \psi_j \rangle &= (1/4\pi^2) \int_0^{2\pi} \int_0^{2\pi} F \sin \psi_j d\psi_1 d\psi_2. \end{aligned} \quad (30)$$

The function $F(t, x, \dot{x}, \ddot{x})$ that appears on the right side of the system of differential equations in (21) to (24) was defined earlier in (8). The quantities x, \dot{x}, \ddot{x} appearing in (8) are next replaced by the trigonometric expressions stated in (12) and (14) to (16). This implies that F becomes a

function of t, ψ_1 and ψ_2 . Thus, it becomes convenient to express F as a double Fourier series as follows:

$$\begin{aligned} F(t, x, \dot{x}, \ddot{x}, \ddot{\ddot{x}}) &= F_*(t, \psi_1, \psi_2) \\ &= \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} [F_{mn}^{(1)}(t) \cos(m\psi_1) + F_{mn}^{(2)}(t) \sin(m\psi_1)] \times \\ &\quad \times [F_{mn}^{(3)}(t) \cos(n\psi_2) + F_{mn}^{(4)}(t) \sin(n\psi_2)] , \end{aligned} \quad (31)$$

where the coefficients $F_{mn}^{(j)}(t)$, $j = 1(1)4$, are known functions of time t . In expanded form, the first few and most important terms of (31) are

$$F_*(t, \psi_1, \psi_2) = F_0 + F_1 \cos \psi_1 + E_1 \sin \psi_1 + G_1 \cos \psi_2 + H_1 \sin \psi_2 + \dots , \quad (32)$$

where the explicit forms of the coefficients F_0, F_1, \dots, H_1 are known.

If we use (32) in (30), we find that

$$\begin{aligned} \langle F \cos \psi_1 \rangle &= F_1/2 , & \langle F \sin \psi_1 \rangle &= E_1/2 , \\ \langle F \cos \psi_2 \rangle &= G_1/2 , & \langle F \sin \psi_2 \rangle &= H_1/2 . \end{aligned} \quad (33)$$

Substitution of (33) into (29) yields the simplified or averaged system of equations

$$\begin{aligned} \dot{a}_1 + \gamma_1 a_1 &= (\epsilon/2)(R_{11}F_1 + R_{12}E_1) , \\ \dot{\theta}_1 &= -(\epsilon/2a_1)(R_{11}E_1 - R_{12}F_1) , \\ \dot{a}_2 + \gamma_2 a_2 &= (\epsilon/2)(R_{22}G_1 - R_{12}H_1) , \\ \dot{\theta}_2 &= -(\epsilon/2a_2)(R_{22}H_1 + R_{12}G_1) . \end{aligned} \quad (34)$$

From this point, the objective is to solve the system of ordinary differential equations in (34) subject to the set of initial conditions that has been stated in (2/).

3. LINEAR DIFFERENTIAL EQUATIONS. As a first application of the theory developed in Section 2, we consider the following linear differential equation with variable coefficients:

$$\ddot{x} + \xi(t)\dot{x} + [\omega_1^2 + n(t)]x = \varepsilon Q(t), \quad t_0 < t. \quad (35)$$

The initial conditions are those stated in (6), and $Q(t)$ is assumed to satisfy the differential equation (5) with $\gamma_2 = 0$. Comparing (35) with (4), we find $\gamma_1 = 0$ and

$$\varepsilon f(t, x, \dot{x}) = \xi(t)\dot{x} + n(t)x. \quad (36)$$

In this case, (8) leads to

$$\varepsilon F(t, x, \dot{x}, \ddot{x}, \ddot{\ddot{x}}) = h_1(t)x + h_2(t)\dot{x} + h_3(t)\ddot{x} + h_4(t)\ddot{\ddot{x}}, \quad (37)$$

where

$$\begin{aligned} h_1(t) &= \ddot{n}(t) + \omega_2^2 n(t), & h_2(t) &= \ddot{\xi}(t) + \omega_2^2 \xi(t) + 2\dot{n}(t), \\ h_3(t) &= n(t) + 2\dot{\xi}(t), & h_4(t) &= \xi(t). \end{aligned} \quad (38)$$

In the event that $\gamma_1 = \gamma_2 = 0$, (25) reduces to

$$R_{11} = \frac{1}{\omega_1(\omega_1^2 - \omega_2^2)}, \quad R_{12} = 0, \quad R_{22} = \frac{-1}{\omega_2(\omega_1^2 - \omega_2^2)}$$

Consequently, the averaged differential equations in (34) become

$$\begin{aligned}
\dot{a}_1 &= \epsilon R_{11} F_1 / 2 & \dot{\theta}_1 &= -\epsilon R_{11} E_1 / 2 a_1 \\
\dot{a}_2 &= \epsilon R_{22} G_1 / 2, & \dot{\theta} &= -\epsilon R_{22} H_1 / 2 a_2
\end{aligned}
\tag{39}$$

where, in view of (32), (37) and (38),

$$\begin{aligned}
F_1 &= (\omega_1 / \epsilon) [\ddot{\xi}(t) + (\omega_2^2 - \omega_1^2) \xi(t) + 2\dot{\eta}(t)] a_1, \\
E_1 &= (1/\epsilon) [\ddot{\eta}(t) + (\omega_2^2 - \omega_1^2) \eta(t) - 2\omega_1^2 \dot{\xi}(t)] a_1, \\
G_1 &= (\omega_2 / \epsilon) [\ddot{\xi}(t) + 2\dot{\eta}(t)] a_2 \\
H_1 &= (1/\epsilon) [\ddot{\eta}(t) - 2\omega_2^2 \dot{\xi}(t)] a_2.
\end{aligned}
\tag{40}$$

When (40) is substituted into (39), the resulting system of first order differential equations is found to be

$$\begin{aligned}
\dot{a}_1 &= \frac{a_1}{2(\omega_1^2 - \omega_2^2)} [\ddot{\xi}(t) + 2\dot{\eta}(t) + (\omega_2^2 - \omega_1^2) \xi(t)], \\
\dot{\theta}_1 &= \frac{-1}{2\omega_1(\omega_1^2 - \omega_2^2)} [\ddot{\eta}(t) - 2\omega_1^2 \dot{\xi}(t)] + \frac{1}{2\omega_1} \eta(t), \\
\dot{a}_2 &= \frac{a_2}{2(\omega_2^2 - \omega_1^2)} [\ddot{\xi}(t) + 2\dot{\eta}(t)], \\
\dot{\theta}_2 &= \frac{-1}{2\omega_2(\omega_2^2 - \omega_1^2)} [\ddot{\eta}(t) - 2\omega_2^2 \dot{\xi}(t)].
\end{aligned}
\tag{41}$$

The system of differential equations in (41) is a linear system of elementary uncoupled differential equations with coefficients that vary with time t .

In view of (6), (10), (11) and (36), the initial conditions associated with the system (41) are the expressions given in (28) with

$$\begin{aligned}
p_0 &= \epsilon Q(t_0) - x_0[\eta(t_0) + \omega_1^2] - v_0 \xi(t_0), \\
q_0 &= \epsilon Q(t_0) - x_0 \dot{\eta}(t_0) - v_0[\eta(t_0) + \omega_1^2 + \dot{\xi}(t_0)] - p_0 \xi(t_0).
\end{aligned}
\tag{42}$$

The differential equations in (41) are easily integrated to yield

$$\begin{aligned}
 a_1(t) &= a_1(t_0) \exp \left\{ \frac{\dot{\xi}(t) + 2\eta(t) - \dot{\xi}(t_0) - 2\eta(t_0)}{2(\omega_1^2 - \omega_2^2)} - \frac{1}{2} \int_{t_0}^t \xi(\tau) d\tau \right\}, \\
 \theta_1(t) &= \frac{1}{2\omega_1} \int_{t_0}^t \eta(\tau) d\tau - \frac{1}{2\omega_1(\omega_1^2 - \omega_2^2)} [\dot{\eta}(t) - 2\omega_1^2 \xi(t) - \dot{\eta}(t_0) + \\
 &\quad + 2\omega_1^2 \xi(t_0)] + \theta_1(t_0), \\
 a_2(t) &= a_2(t_0) \exp \left\{ \frac{\dot{\xi}(t) + 2\eta(t) - \dot{\xi}(t_0) - 2\eta(t_0)}{2(\omega_2^2 - \omega_1^2)} \right\},
 \end{aligned}
 \tag{43}$$

$$\theta_2(t) = \frac{-1}{2\omega_2(\omega_2^2 - \omega_1^2)} [\dot{\eta}(t) - 2\omega_2^2 \xi(t) - \dot{\eta}(t_0) + 2\omega_2^2 \xi(t_0)] + \theta_2(t_0).$$

The quantities $a_j(t_0)$ and $\theta_j(t_0)$, $j = 1, 2$, are to be determined from (28)₂ and (42).

Example 1. We consider the following nonhomogeneous Mathieu's equation:

$$\ddot{x} + (\omega_1^2 + 4b \cos 2t)x = \varepsilon Q_0 \sin \omega_2 t, \quad 0 < t. \tag{44}$$

A comparison of (3b) and (44) reveals that

$$\xi(t) = 0, \quad \eta(t) = 4b \cos 2t, \quad t_0 = 0, \quad Q(t) = Q_0 \sin \omega_2 t. \tag{45}$$

Substituting (45) into (43), we find

$$a_1(t) = a_1(0) \exp \left[\frac{4b(\cos 2t - 1)}{\omega_1^2 - \omega_2^2} \right],$$

$$\theta_1(t) = \frac{b}{\omega_1(\omega_1^2 - \omega_2^2)}(\omega_1^2 - \omega_2^2 + 4)\sin 2t + \theta_1(t_0) ,$$

$$a_2(t) = a_2(t_0)\exp\left[\frac{4b(1 - \cos 2t)}{\omega_1^2 - \omega_2^2}\right] ,$$

(46)

$$\theta_2(t) = \frac{4b}{\omega_2(\omega_2^2 - \omega_1^2)} \sin 2t + \theta_2(t_0) .$$

In the present case, (42) reduces to

$$p_0 = -x_0(\omega_1^2 + 4b), \quad q_0 = \varepsilon\omega_2 Q_0 - v_0(\omega_1^2 + 4b) ,$$

by virtue of (45). For the initial conditions $x_0 = v_0 = 0$, we obtain $p_0 = 0$ and $q_0 = \varepsilon\omega_2 Q_0$. Hence, (28a) leads to

$$a_1(0) = \frac{\varepsilon\omega_2 Q_0}{\omega_1(\omega_2^2 - \omega_1^2)} , \quad \theta_1(0) = 0 ,$$

$$a_2(0) = \frac{\varepsilon Q_0}{\omega_1^2 - \omega_2^2} , \quad \theta_2(0) = 0 .$$

Consequently, (46) can now be expressed as

$$a_1(t) = \frac{\varepsilon\omega_2 Q_0}{\omega_1(\omega_2^2 - \omega_1^2)} \exp\left[\frac{4b(\cos 2t - 1)}{\omega_1^2 - \omega_2^2}\right] ,$$

$$\theta_1(t) = \frac{b}{\omega_1(\omega_1^2 - \omega_2^2)}(\omega_1^2 - \omega_2^2 + 4)\sin 2t ,$$

$$a_2(t) = \frac{\varepsilon Q_0}{\omega_1^2 - \omega_2^2} \exp\left[\frac{4b(1 - \cos 2t)}{\omega_1^2 - \omega_2^2}\right] ,$$

(47)

$$\theta_2(t) = \frac{4b}{\omega_2(\omega_2^2 - \omega_1^2)} \sin 2t .$$

Finally, substitution of (47) into (12) leads to the following approximation for the solution of (44) subject to the homogeneous initial conditions $\dot{x}(0) = x(0) = 0$.

$$x(t) \approx \frac{\epsilon Q_0}{\omega_2^2 - \omega_1^2} \left\{ (\omega_2/\omega_1) \exp\left[\frac{-4b(1 - \cos 2t)}{\omega_1^2 - \omega_2^2}\right] \sin\left[\omega_1 t + \frac{b(\omega_1^2 - \omega_2^2 + 4)}{\omega_1(\omega_1^2 - \omega_2^2)} \sin 2t\right] - \exp\left[\frac{4b(1 - \cos 2t)}{\omega_1^2 - \omega_2^2}\right] \sin\left[\omega_2 t + \frac{4b}{\omega_2(\omega_2^2 - \omega_1^2)} \sin 2t\right] \right\}. \quad (48)$$

This approximate solution for (44) consists of the sum of two sinusoidally varying terms whose arguments consist of the sum of a linear term and a sine term in t . The magnitude of the phase angle becomes very large as the value of ω_2 tends toward ω_1 . In addition, the primary sinusoidal terms are multiplied by exponential functions whose arguments depend upon $\cos 2t$. These factors cannot give rise to an unlimited growth in time of the amplitude of motion. Since the leading coefficient in (48) contains the term $\omega_2^2 - \omega_1^2$ in the denominator, the value of x will become very large whenever the value of ω_2 is close to that of ω_1 .

To assess the accuracy of the approximation for $x(t)$ given in (48), we have evaluated this approximate solution and have solved (44) subject to the initial conditions $\dot{x}(0) = x(0) = 0$ numerically by means of the Runge-Kutta method (hereinafter called the "exact" solution). The numerical results obtained in this manner have been plotted on the same set of axes (x versus t) for purposes of comparison. For the choice of parameters $\omega_1 = 1.5$, $\omega_2 = 6$, $b = 0.05$, $Q_0 = 1$ and $\epsilon = 0.1$, the exact and approximate curves for $x(t)$ versus t have been plotted in Figure 1 on the interval $0 \leq t \leq 10$. In general, the quantitative agreement between the approximate and exact curves is quite good over the entire interval shown. If the values of all the preceding parameters are held constant while the value of ω_2 is diminished from $\omega_2 = 6$ to $\omega_2 = 4$, then numerical calculations yield the plot of $x(t)$ versus t shown in Figure 2. Once again, good agreement between the exact and approximate curves is very good.

The approximation in (48) leads to somewhat less satisfactory results when the frequency parameters ω_1 and ω_2 were assigned the values of $\omega_1 = 1.1$ and $\omega_2 = 3$. This is not surprising since the nature of the solution of the homogeneous form of Mathieu's equation is known to change when the value of ω_1 is close to unity--see, for example, Nayfeh [11], pp. 234 to 254.

Example 2. Consider next the non-homogeneous form of Bessel's equation, namely,

$$\ddot{x} + (1/t)\dot{x} + (\omega_1^2 - v^2/t^2)x = \epsilon Q(t), \quad 0 < t_0 < t, \quad (49)$$

where $Q(t)$ is assumed to satisfy the differential equation

$$\ddot{Q} + \omega_2^2 Q = 0. \quad (49a)$$

Upon referring to (35), we see that for (49)

$$\xi(t) = 1/t, \quad \eta(t) = -v^2/t^2. \quad (50)$$

Consequently, (43) leads to

$$\begin{aligned} a_1(t) &= a_1(t_0)(t_0/t)^{1/2} \exp\left[\frac{1+2v^2}{2(\omega_1^2 - \omega_2^2)}\left(\frac{1}{t_0^2} - \frac{1}{t^2}\right)\right], \\ \theta_1(t) &= \left[\frac{(v^2 + 2)\omega_1^2 - (v\omega_2)^2}{2\omega_1(\omega_1^2 - \omega_2^2)}\right]\left(\frac{1}{t} - \frac{1}{t_0}\right) - \frac{v^2}{\omega_1(\omega_1^2 - \omega_2^2)}\left(\frac{1}{t^3} - \frac{1}{t_0^3}\right) + \theta_1(t_0), \\ a_2(t) &= a_2(t_0) \exp\left[\frac{1+2v^2}{2(\omega_2^2 - \omega_1^2)}\left(\frac{1}{t_0^2} - \frac{1}{t^2}\right)\right], \\ \theta_2(t) &= \frac{-1}{\omega_2(\omega_2^2 - \omega_1^2)}\left[v^2\left(\frac{1}{t^3} - \frac{1}{t_0^3}\right) - \omega_2^2\left(\frac{1}{t} - \frac{1}{t_0}\right)\right] + \theta_2(t_0), \end{aligned} \quad (51)$$

where the quantities $a_j(t_0)$ and $\theta_j(t_0)$, $j = 1, 2$, are to be determined from $(28)_2$.

In case of $x_0 = v_0 = 0$ and $Q(t) = Q_0 \sin \omega_2 t$, we find from (49) that

$$p_0 = \varepsilon Q_0 \sin \omega_2 t_0, \quad q_0 = \varepsilon Q_0 [\omega_2 \cos(\omega_2 t_0) - (1/t_0) \sin(\omega_2 t_0)].$$

Therefore, $(28)_2$ leads to

$$\begin{aligned} a_1(t_0) &= [(q_0/\omega_1)^2 + p_0^2]^{1/2} / |\omega_2^2 - \omega_1^2|, \\ \theta_1(t_0) &= \tan^{-1} \left[\frac{\omega_1 \sin(\omega_2 t_0)}{\omega_2 \cos(\omega_2 t_0) - (1/t_0) \sin(\omega_2 t_0)} \right] - \omega_1 t_0, \\ a_2(t_0) &= -[(q_0/\omega_2)^2 + p_0^2]^{1/2} / |\omega_2^2 - \omega_1^2|, \\ \theta_2(t_0) &= \tan^{-1} \left[\frac{\omega_2 \sin(\omega_2 t_0)}{\omega_2 \cos(\omega_2 t_0) - (1/t_0) \sin(\omega_2 t_0)} \right] - \omega_2 t_0. \end{aligned} \quad (52)$$

Therefore, an approximation for $x(t)$ is

$$\begin{aligned} x(t) &\approx a_1(t_0) (t_0/t)^{1/2} \exp \left[\frac{1 + 2v^2}{2(\omega_1^2 - \omega_2^2)} \left(\frac{1}{t_0^2} - \frac{1}{t^2} \right) \right] \sin[\omega_1 t + \\ &\quad + \frac{[(2 + v^2)\omega_1^2 - v^2\omega_2^2]}{2\omega_1(\omega_1^2 - \omega_2^2)} \left(\frac{1}{t} - \frac{1}{t_0} \right) + \frac{v^2}{\omega_1(\omega_1^2 - \omega_2^2)} \left(\frac{1}{t_0^3} - \frac{1}{t^3} \right) + \theta_1(t_0)] + \\ &\quad + a_2(t_0) \exp \left[\frac{1 + 2v^2}{2(\omega_2^2 - \omega_1^2)} \left(\frac{1}{t_0^2} - \frac{1}{t^2} \right) \right] \sin[\omega_2 t + \\ &\quad + \frac{1}{\omega_2(\omega_2^2 - \omega_1^2)} [v^2 \left(\frac{1}{t_0^3} - \frac{1}{t^3} \right) + \omega_2^2 \left(\frac{1}{t} - \frac{1}{t_0} \right)] + \theta_2(t_0)]. \end{aligned} \quad (53)$$

As t becomes very large, the coefficient of the first sine term in (53) tends to zero. Consequently, as $t \rightarrow \infty$, (53) reduces to

$$x(t) \approx a_2(t_0) \exp\left[\frac{1+2v^2}{2t_0^2(\omega_2^2 - \omega_1^2)}\right] \sin[\omega_2(t - t_0) - \frac{\omega_2}{t_0(\omega_2^2 - \omega_1^2)} + \frac{v^2}{\omega_2 t_0^3(\omega_2^2 - \omega_1^2)} + \tan^{-1}\left[\frac{\omega_2 \sin \omega_2 t_0}{\omega_2 \cos \omega_2 t_0 - (1/t_0) \sin \omega_2 t_0}\right]]. \quad (54)$$

As in the preceding example, numerical calculations have been performed to solve

$$\ddot{x} + (1/t)\dot{x} + (\omega_1^2 - v^2/t^2)x = \epsilon Q_0 \sin \omega_2 t, \quad 0 < t_0 < t, \quad (55)$$

subject to the initial conditions $x(t_0) = \dot{x}(t_0) = 0$. Some typical plots are shown in Figures 3 and 4, for which the following numerical values for the parameters were used: $t_0 = 1$, $\omega_1 = 1$, $Q_0 = 1$ and $\epsilon = 0.1$. The value $\omega_2 = 6$, $v = 1/4$ and $\omega_2 = 4$, $v = 1$ were used in the plotting of Figures 3 and 4, respectively. Figure 3 shows that there is very good agreement between the approximate and exact curves over the entire time interval considered, namely $1 \leq t \leq 10$. The approximation proves to be slightly less precise in Figure 4, where ω_2 has a smaller value. The approximate curve, nonetheless, respects the qualitative features of the exact curve reasonably well.

Example 3. The nonhomogeneous form of Hermite's differential equation (see Sneddon [12], p. 152) is

$$\ddot{x} - 2t\dot{x} + 2vx = \epsilon Q(t), \quad 0 < t, \quad (56)$$

where $Q(t)$ satisfies (b) when $\gamma_2 = 0$. Since (56) is a particular case of (35), we have

$$\omega_1^2 = 2v, \quad \xi(t) = -2t, \quad \eta(t) = 0, \quad t_0 = 0. \quad (57)$$

When the expressions in (57) are inserted into (43), the results are found to be

$$\begin{aligned} a_1(t) &= a_1(0)e^{t^2/2}, & \theta_1(t) &= \theta_1(0) - \frac{2\omega_1 t}{\omega_1^2 - \omega_2^2}, \\ a_2(t) &= a_2(0), & \theta_2(t) &= \theta_2(0) - \frac{2\omega_2 t}{\omega_2^2 - \omega_1^2}, \end{aligned} \quad (58)$$

where for $x_0 = v_0 = 0$ and $Q(t) = Q_0 \sin \omega_2 t$,

$$\begin{aligned} a_1(0) &= \frac{\epsilon \omega_2 Q_0}{\omega_1 |\omega_2^2 - \omega_1^2|}, & a_2(0) &= \frac{-\epsilon Q_0}{|\omega_2^2 - \omega_1^2|}, \\ \theta_1(0) &= \theta_2(0) = 0. \end{aligned} \quad (59)$$

When (58) and (59) are substituted into (12), the result is the following approximation for $x(t)$:

$$\begin{aligned} x(t) &\approx \frac{\epsilon Q_0}{|\omega_2^2 - \omega_1^2|} \left\{ (\omega_2/\omega_1) e^{t^2/2} \sin[\omega_1 t (1 - \frac{2}{\omega_1^2 - \omega_2^2})] - \right. \\ &\quad \left. - \sin[\omega_2 t (1 - \frac{2}{\omega_2^2 - \omega_1^2})] \right\}. \end{aligned} \quad (60)$$

Numerical results based upon the approximate solution (60) and the "exact" Runge-Kutta solution have been plotted on the interval $0 \leq t \leq 1.5$ in Figures 5 and 6, where the set of parameters $\omega_2 = 6$, $\nu = 3$ and $\omega_2 = 10$, $\nu = 6$, respectively, have been used with $Q_0 = 1$, $\epsilon = 0.1$. The agreement between the approximate and the "exact" solutions is generally rather good for the time interval considered.

4. A NONLINEAR DIFFERENTIAL EQUATION. In the preceding section, we showed that the general theory developed in Section 2 can be applied effectively to several nonhomogeneous linear differential equations with variable coefficients for the purpose of generating analytical expressions that represent approximate solutions. It is next of interest to examine a forced motion problem in the theory of nonlinear oscillations. In particular, we consider a form of Duffing's equation that includes the effect of linear damping, i.e.,

$$\ddot{x} + 2\gamma\dot{x} + \omega^2x + \epsilon\alpha x^3 = \epsilon Q(t), \quad 0 < t, \quad (61)$$

where $Q(t)$ satisfies (b). Comparing (4) and (61), we write

$$\gamma_1 = \gamma, \quad \omega_1 = \omega, \quad f(t, x, \dot{x}) = \alpha x^3, \quad (62)$$

where $\alpha = 1$ for a hard spring and $\alpha = -1$ for a soft spring. Using (62) and (8), we can easily show that

$$\begin{aligned} F &= \omega_2^2 f + 2\gamma_2 \frac{\partial f}{\partial \dot{x}} \dot{x} + \frac{\partial f}{\partial x} \ddot{x} + \frac{\partial^2 f}{\partial x^2} \dot{x}^2 \\ &= \alpha \omega_2^2 x^3 + 6\alpha \gamma_2 x^2 \dot{x} + 3\alpha x^2 \ddot{x} + 6\alpha x \dot{x}^2. \end{aligned} \quad (63)$$

If we now use (12), (14) and (15) with (63), we can verify that

$$\begin{aligned} \alpha F &= A_1 \sin^3 \psi_1 + A_2 \sin^2 \psi_1 \sin \psi_2 + A_3 \sin \psi_1 \sin^2 \psi_2 + \\ &+ A_4 \sin^3 \psi_2 + A_5 \sin^2 \psi_1 \cos \psi_1 + A_6 \sin^2 \psi_1 \cos \psi_2 + \\ &+ A_7 \sin \psi_1 \cos \psi_1 \sin \psi_2 + A_8 \sin \psi_1 \sin \psi_2 \cos \psi_2 + \\ &+ A_9 \cos \psi_1 \sin^2 \psi_2 + A_{10} \sin^2 \psi_2 \cos \psi_2 + A_{11} \sin \psi_1 \cos^2 \psi_1 + \\ &+ A_{12} \sin \psi_1 \cos^2 \psi_2 + A_{13} \sin \psi_1 \cos \psi_1 \cos \psi_2 + A_{14} \cos^2 \psi_1 \sin \psi_2 + \\ &+ A_{15} \sin \psi_2 \cos^2 \psi_2 + A_{16} \cos \psi_1 \sin \psi_2 \cos \psi_2, \end{aligned} \quad (64)$$

where

$$\begin{aligned}
A_1 &= (\omega_2^2 - 3\omega_1^2 + 12\gamma_1^2 - 6\gamma_1\gamma_2)a_1^3, & A_2 &= 6(3\gamma_1^2 - \omega_1^2)a_1^2a_2, \\
A_3 &= 3(2\gamma_1^2 + 2\gamma_1\gamma_2 + 2\gamma_2^2 - \omega_1^2 - \omega_2^2)a_1a_2^2, & A_4 &= 2(3\gamma_2^2 - \omega_2^2)a_2^3, \\
A_5 &= 6\Omega_1(\gamma_2 - 3\gamma_1)a_1^3, & A_6 &= -12\gamma_1\Omega_2a_1^2a_2, \\
A_7 &= -24\gamma_1\Omega_1a_1^2a_2, & A_8 &= -12(\gamma_1 + \gamma_2)\Omega_2a_1a_2^2, \\
A_9 &= -6\Omega_1(\gamma_1 + \gamma_2)a_1a_2^2, & A_{10} &= -12\gamma_2\Omega_2a_2^3, & A_{11} &= 6\Omega_1^2a_1^3, & (65) \\
A_{12} &= 6\Omega_2^2a_1a_2^2, & A_{13} &= 12\Omega_1\Omega_2a_1^2a_2, & A_{14} &= 6\Omega_1^2a_1^2a_2, \\
A_{15} &= 6\Omega_2a_2^3, & A_{16} &= 12\Omega_1\Omega_2a_1a_2^2.
\end{aligned}$$

If (64) is used with (30) and (33), we find after some manipulation that

$$\begin{aligned}
\langle F\cos\psi_1 \rangle &= (\alpha/8)(A_5 + 2A_9), \\
\langle F\sin\psi_1 \rangle &= (\alpha/8)(3A_1 + 2A_3 + A_{11} + 2A_{12}), & (66) \\
\langle F\cos\psi_2 \rangle &= (\alpha/8)(2A_6 + A_{10}), \\
\langle F\sin\psi_2 \rangle &= (\alpha/8)(2A_2 + 3A_4 + 2A_{14} + A_{15}),
\end{aligned}$$

where it may be observed that the coefficients A_7 , A_8 , A_{13} and A_{16} defined in (65) do not appear. In view of (66) and the right side of (29), it can be shown that

$$\begin{aligned}
 R_{11} < F \cos \psi_1 > + R_{12} < F \sin \psi_1 > &= H_1 a_1^3 + H_2 a_1 a_2^2, \\
 R_{11} < F \sin \psi_1 > - R_{12} < F \cos \psi_1 > &= H_3 a_1^3 + H_4 a_1 a_2^2, \\
 R_{22} < F \cos \psi_2 > - R_{12} < F \sin \psi_2 > &= -H_5 a_1^2 a_2 - H_6 a_2^3, \\
 R_{22} < F \sin \psi_2 > + R_{12} < F \cos \psi_2 > &= H_7 a_1^2 a_2 + H_8 a_2^3,
 \end{aligned} \tag{67}$$

where

$$\begin{aligned}
 H_1 &= (3\alpha\gamma_1/2R)[\omega_2^2 - \omega_1^2 - 2(\gamma_2 - \gamma_1)^2], \quad H_2 = (3\alpha\gamma_2/R)(\omega_2^2 - \omega_1^2), \\
 H_3 &= -(3\alpha/8\Omega_1 R)\{(\omega_2^2 - \omega_1^2)^2 + 4\gamma_1(3\gamma_1 - 2\gamma_2)(\omega_2^2 - \omega_1^2) + \\
 &\quad + 8\gamma_1^2(\gamma_2 - \gamma_1)^2 + 4\omega_1^2(\gamma_2 - \gamma_1)(\gamma_2 - 3\gamma_1)\}, \\
 H_4 &= -(3\alpha/4\Omega_1 R)[(\omega_2^2 - \omega_1^2)^2 + 4\gamma_1^2(\omega_2^2 - \omega_1^2) - 4\omega_1^2(\gamma_2^2 - \gamma_1^2)], \\
 H_5 &= (3\alpha\gamma_1/R)[\omega_2^2 - \omega_1^2 - 2(\gamma_2 - \gamma_1)^2], \quad H_6 = (3\alpha\gamma_2/2R)(\omega_2^2 - \omega_1^2), \\
 H_7 &= (3\alpha\gamma_1/\Omega_2 R)[(3\gamma_1 - 2\gamma_2)\omega_2^2 - \gamma_1\omega_1^2 + 2\gamma_2(\gamma_2 - \gamma_1)^2], \\
 H_8 &= (3\alpha\gamma_2/2\Omega_2 R)[(2\gamma_1 - \gamma_2)\omega_2^2 - \gamma_2\omega_1^2].
 \end{aligned} \tag{68}$$

As a consequence of (67), (29) can now be expressed as

$$\begin{aligned}
 \dot{a}_1 + \gamma_1 a_1 &= \epsilon(H_1 a_1^3 + H_2 a_1 a_2^2), \\
 \dot{\theta}_1 &= -\epsilon(H_3 a_1^2 + H_4 a_2^2), \\
 \dot{a}_2 + \gamma_2 a_2 &= -2\epsilon(H_1 a_1^2 a_2 + H_2 a_2^3), \\
 \dot{\theta}_2 &= -\epsilon(H_7 a_1^2 + H_8 a_2^2).
 \end{aligned} \tag{69}$$

Example 4. Suppose that $\gamma_2 = 0$, so that $Q(t)$ is a solution of (49a). it now follows from (68) that

$$\begin{aligned} H_1 &= (3\alpha\gamma_1/2R)(\omega_2^2 - \omega_1^2 - 2\gamma_1^2), & H_2 &= 0, \\ H_3 &= -(3\alpha/8\Omega_1 R)[(\omega_2^2 - \omega_1^2)^2 + 12\gamma_1^2\omega_2^2 + 8\gamma_1^4], & (70) \\ H_4 &= -3\alpha/4\Omega_1, & H_7 &= (3\alpha\gamma_1^2/\omega_2 R)(3\omega_2^2 - \omega_1^2), & H_8 &= 0, \end{aligned}$$

where, in view of (25),

$$R = (\omega_2^2 - \omega_1^2)^2 + (2\gamma_1\omega_2)^2. \quad (71)$$

In view of (70), the averaged equations of motion (69) reduce to

$$\dot{a}_1 + \gamma_1 a_1 = \epsilon H_1 a_1^3, \quad (72)$$

$$\dot{\theta}_1 = -\epsilon(H_3 a_1^2 + H_4 a_2^2), \quad (73)$$

$$\dot{a}_2 = -2\epsilon H_1 a_1^2 a_2, \quad (74)$$

$$\dot{\theta}_2 = -\epsilon H_7 a_1^2. \quad (75)$$

The differential equation (72) is uncoupled but nonlinear. It can be solved by elementary methods. Indeed, we can write

$$\frac{da_1}{a_1(a_1^2 - c^2)} = \epsilon H_1 dt, \quad (76)$$

where we assume that $\epsilon > 0, \alpha > 0$ and $\omega_2^2 > \omega_1^2 + 2\gamma_1^2$, so that $H_1 > 0$ and

$$c^2 = \gamma_1/\epsilon H_1 > 0. \quad (77)$$

The integral of (76) is easily verified to be

$$a_1(t) = \frac{ce^{-\gamma_1 t}}{(r + e^{-2\gamma_1 t})^{1/2}} \quad (78)$$

where

$$r = [c/a_1(0)]^2 - 1. \quad (79)$$

Since the explicit form of $a_1(t)$ is now available in (78), we can determine $a_2(t)$ through an integration of (74). The result is

$$a_2(t) = \sigma(r + e^{-2\gamma_1 t}), \quad (80)$$

where

$$\sigma = a_2(0)[a_1(0)/c]^2. \quad (81)$$

Finally, using the expressions for $a_1(t)$ and $a_2(t)$ in (78) and (80), we derive the following phase angle expressions from (73) and (75):

$$\begin{aligned} \theta_1(t) = & (H_3/2H_1) \ln \left[\frac{r + e^{-2\gamma_1 t}}{r + 1} \right] - \frac{\epsilon H_4 \sigma^2}{4\gamma_1} [(2r + 1)^2 + \gamma_1 t (2r)^2 - \\ & - (2r + e^{-2\gamma_1 t})^2] + \theta_1(0) \end{aligned} \quad (82)$$

and

$$\theta_2(t) = (H_7/2H_1) \ln \left[\frac{r + e^{-2\gamma_1 t}}{r + 1} \right] + \theta_2(0). \quad (83)$$

Expressions for the initial values $a_j(0)$ and $\theta_j(0)$, $j = 1, 2$, must now be determined. When $\gamma_2 = 0$, (28) yields

$$\begin{aligned} W_1 &= \gamma_1(x_0\omega_2^2 + p_0)(\omega_2^2 - 3\omega_1^2 + 4\gamma_1^2) + (v_0\omega_2^2 + q_0)(\omega_2^2 - \omega_1^2 + 2\gamma_1^2), \\ W_2 &= (x_0\omega_2^2 + p_0)(\omega_2^2 - \omega_1^2 + 4\gamma_1^2) + 2\gamma_1(v_0\omega_2^2 + q_0), \\ Z_1 &= 2\gamma_1\omega_2^2(x_0\omega_2^2 + p_0) + v_0[\omega_1^2(\omega_1^2 - \omega_2^2) + (2\gamma_1\omega_2)^2] + q_0(\omega_1^2 - \omega_2^2), \\ Z_2 &= x_0\omega_1^2(\omega_1^2 - \omega_2^2) - p_0(\omega_2^2 - \omega_1^2 + 4\gamma_1^2) - 2\gamma_1(v_0\omega_2^2 + q_0). \end{aligned} \quad (84)$$

Furthermore, from (10) and (11), we find

$$\begin{aligned} p_0 &= \epsilon Q(0) - \omega_1^2 x_0 - 2\gamma_1 v_0 - \epsilon \alpha x_0^3, \\ q_0 &= \epsilon \dot{Q}(0) - \omega_1^2 v_0 - 2\gamma_1 p_0 - 3\epsilon \alpha v_0 x_0^2, \end{aligned} \quad (85)$$

since $f(t, x, \dot{x}) = \alpha x^3$. In the case of $Q(t) = Q_0 \sin(\omega_2 t)$ and $x_0 = v_0 = 0$, we have from (85)

$$p_0 = 0, \quad q_0 = \epsilon \omega_2 Q_0, \quad (86)$$

whereas (84) yields

$$\begin{aligned} W_1 &= \epsilon \omega_2 Q_0 (\omega_2^2 - \omega_1^2 + 2\gamma_1^2), & W_2 &= 2\epsilon \gamma_1 \omega_2 Q_0, \\ Z_1 &= q_0 (\omega_1^2 - \omega_2^2), & Z_2 &= -2\gamma_1 q_0. \end{aligned}$$

Consequently, (27) leads to the expressions

$$\begin{aligned} a_1(0) &= (\epsilon \omega_2 Q_0 / \Omega_1 R) [(\omega_2^2 - \omega_1^2)^2 + 2\gamma_1^2(\omega_1^2 + \omega_2^2)]^{1/2}, \\ a_2(0) &= -\epsilon Q_0 / R^{1/2}, \end{aligned} \quad (87)$$

$$\theta_1(0) = \tan^{-1} \left[\frac{2\gamma_1 \Omega_1}{\omega_2^2 - \omega_1^2 + 2\gamma_1^2} \right], \quad \theta_2(0) = \tan^{-1} \left[\frac{2\gamma_1 \omega_2}{\omega_2^2 - \omega_1^2} \right],$$

where R was given earlier in (71).

Substituting (78) and (80) into (12), we find

$$x(t) \approx \frac{ce^{-\gamma_1 t}}{(r + e^{-2\gamma_1 t})^{1/2}} \sin[\Omega_1 t + \theta_1(t)] + a_2(0) \frac{a_1^2(0)}{c^2} (r + e^{-2\gamma_1 t}) \sin[\omega_2 t + \theta_2(t)], \quad (88)$$

where $\theta_j(t)$, $j = 1, 2$, are defined in (82) and (83), with the initial values being given in (87).

It is easy to see that the function $a_1(t)$ in (78) approaches zero as t tends to infinity, whereas according to (80) and (83)

$$a_2(t) \sim a_2(0) [1 - a_1^2(0)/c^2],$$

$$\theta_2(t) \sim (H_7/2H_1) \ln\left(\frac{r}{r+1}\right) + \tan^{-1} \left[\frac{2\gamma_1 \omega_2}{\omega_2^2 - \omega_1^2} \right]$$

as $t \rightarrow \infty$. Consequently, in the steady state, we have

$$x(t) \sim a_2(0) [1 - a_1^2(0)/c^2] \sin\{\omega_2 t + (H_7/2H_1) \ln\left(\frac{r}{r+1}\right) + \tan^{-1} \left[\frac{2\gamma_1 \omega_2}{\omega_2^2 - \omega_1^2} \right]\}. \quad (89)$$

In the case of $\alpha = 1$, $\omega_1 = 1$, $\omega_2 = 6$, $Q_0 = 1$ and $\varepsilon = 0.1$, plots of the variation of $x(t)$ with t as determined from (88) are shown in Figures 7 and 8 for $\gamma_1 = 0.25$ and $\gamma_1 = 0.75$, respectively. The exact and approximate curves in these figures are in excellent agreement on the time intervals considered. In Figure 8, the higher coefficient of damping ($\gamma_1 = 0.75$) causes a rapid decay of the transient component of the motion, and the steady state behavior as described by (89) becomes evident after the first two cycles.

If the nonlinearity of the system is characterized by a soft spring ($\alpha = -1$), the form of the solution just presented must be modified slightly. We write

$$\beta = -H_1 = (3\gamma_1/2R)(\omega_2^2 - \omega_1^2 - 2\gamma_1^2) ,$$

$$H_3 = (3/8\Omega_1 R)[(\omega_2^2 - \omega_1^2)^2 + 3(2\gamma_1\omega_2)^2 + 8\gamma_1^4] , \quad (90)$$

$$H_4 = -3/4\Omega_1 , \quad H_7 = -(3\gamma_1^2/\omega_2 R)(3\omega_2^2 - \omega_1^2) .$$

Equations (73) and (75) remain unchanged, but it is advantageous to replace (72) and (74) with

$$\dot{a}_1 + \gamma_1 a_1 = -\epsilon\beta a_1^3 , \quad (91)$$

and

$$\dot{a}_2 = 2\epsilon\beta a_1^2 a_2 , \quad (92)$$

respectively.

Integration of (91) leads to

$$a_1(t) = \frac{ve^{-\gamma_1 t}}{(M^2 - e^{-2\gamma_1 t})^{1/2}} , \quad (93)$$

where

$$v^2 = \gamma_1/\epsilon\beta = 2R/3\epsilon(\omega_2^2 - \omega_1^2 - 2\gamma_1^2) , \quad M^2 = 1 + [v/a_1(0)]^2 . \quad (94)$$

The integrations of (92), (73) and (75) are elementary. The results are

$$a_2(t) = a_2(0)a_1^2(0)(M^2 - e^{-2\gamma_1 t})/v^2 , \quad (95)$$

$$\theta_1(t) = \theta_1(0) - (\epsilon/4\gamma_1)[2H_3 v^2 \ln \frac{M^2 - e^{-2\gamma_1 t}}{M^2 - 1}] +$$

$$+ H_4 [a_1^2(0) a_2(0) / v^2] [4\gamma_1 M^4 t + (1 - 2M^2)^2 - (2M^2 - e^{-2\gamma_1 t})^2] , \quad (96)$$

$$\theta_2(t) = \theta_2(0) - (\epsilon H_7 v^2 / 2\gamma_1) \ln \left[\frac{M^2 - e^{-2\gamma_1 t}}{M^2 - 1} \right] , \quad (97)$$

where the values of $a_j(0)$ and $\theta_j(0)$, $j = 1, 2$, can be determined from (87).

Therefore, by virtue of (12), (93) and (95) to (97), an approximate solution of

$$\ddot{x} + 2\gamma_1 \dot{x} + \omega_1^2 x - \epsilon x^3 = \epsilon Q_0 \sin \omega_2 t , \quad (98)$$

$$x(0) = \dot{x}(0) = 0 ,$$

is

$$x(t) \approx \frac{v e^{-\gamma_1 t}}{(M^2 - e^{-2\gamma_1 t})^{1/2}} \sin[\Omega_1 t + \theta_1(t)] +$$

$$+ a_2(0) \frac{a_1^2(0)}{v^2} (M^2 - e^{-2\gamma_1 t}) \sin[\omega_2 t + \theta_2(t)] . \quad (99)$$

As $t \rightarrow \infty$, the solution in (99) tends to the steady state solution

$$x(t) \sim a_2(0)(1 + \xi) \sin(\omega_2 t + \psi_2) , \quad (100)$$

where

$$\xi = a_1^2(0) / v^2 , \quad \psi_2 = \tan^{-1} \left(\frac{2\gamma_1 \omega_2}{\omega_2^2 - \omega_1^2} \right) - (\epsilon H_7 v^2 / 2\gamma_1) \ln(1 + \xi) .$$

To assess the quality of the approximation in (99), we have also solved numerically the initial value problem stated in (98). These "exact" and approximate solutions have been plotted in Figures 9 and 10 for $\omega_1 = 1$, $\omega_2 = 4$, $Q_0 = 1$, and $\epsilon = 0.1$, with $\gamma_1 = 0.25$ and $\gamma_1 = 0.75$, respectively. In both cases, the exact and approximate curves for $x(t)$ versus t are in excellent agreement. The decay of the transient portion of the solution is quite evident. Moreover, the attainment of the steady state motion is clearly visible in Figure 10 since the value of the damping coefficient is three times greater than the value used in the plotting of Figure 9.

5. SUMMARY AND CONCLUSIONS. A method that can be applied for the determination of approximate analytical solutions of the class of nonhomogeneous linear and nonlinear differential equations of second order stated in (4) has been developed. It is hypothesized that the forcing function $Q(t)$ in (4) satisfies the linear second order differential equation with constant coefficients shown in (5). The class of physical problems considered is characteristic of the linear and nonlinear theories of oscillatory systems. The present method represents an extension of Burnelle's method of averaging [3], which specifically accounts for the presence of linear but subcritical damping in the system. Burnelle's technique, which is applicable to homogeneous differential equations, may be viewed as a generalization of the Krylov-Bogoliubov method of averaging [1], in which only small linear damping is considered. The linear damping term is understood to be a component of $f(t, x, \dot{x})$.

In the technique presented here, $x(t)$ is assumed to possess a solution in the form shown in (12), where the amplitudes $a_j(t)$ and the phase of angles $\theta_j(t)$, $j = 1, 2$, are to be determined. The method of variation of parameters is applied, and a system of first order differential equations in the a_j 's and θ_j 's is derived. To simplify this system of differential equations, the averaging operators in (30) are applied to (21) to (24). It may be argued that the right sides of these differential equations are slowly varying functions of time t . This process reduces the complicated system to a much

simpler system that is in many cases tractable by elementary methods. In essence, this has been accomplished by expressing the right sides of (21) to (24) as double Fourier series in ψ_1 and ψ_2 and then approximating the resulting expressions through retention of only the constant term (i.e., the average value of the right side) of the Fourier series. It is this averaging process that introduces the approximation nature into the technique of solution.

The averaged differential equations in (34) were then solved for some specific choices of the function $f(t, x, \dot{x})$ that appears in (4). Firstly, a class of linear differential equations was considered for the form of $f(t, x, \dot{x})$ shown in (36). The integrated forms of $a_j(t)$ and $\theta_j(t)$ are presented in (43). As illustrative examples, nonhomogeneous forms of the differential equations of Hermite, Mathieu and Bessel have been solved approximately by the proposed method. To assess the validity of the analytical approximations, these same initial value problems were solved numerically for several values of the parameters by the Runge-Kutta method, herein called the "exact" method. Plots of $x(t)$ versus t based upon both the analytical approximations and the numerically exact computations were prepared on the same sets of axes. These numerical results, shown in Figures 1 to 6, proved to be in excellent qualitative and relatively good quantitative agreement.

Secondly, Duffing's nonlinear differential equation including the effect of subcritical linear damping has also been solved in an approximate sense by the proposed method. Both hard and soft nonlinear springs have been considered, i.e., $f(t, x, \dot{x}) = \pm x^3$. Numerical calculations revealed that the analytical approximation provides a very accurate representation of the solution of the nonlinear differential equation even into the steady state domain. The exact and approximate curves have been plotted in Figures 7 to 10. The human eye cannot distinguish between them on the scale of the axes employed.

Consequently, the method developed here offers a practical technique for the determination of rather accurate analytical approximate solutions for certain types of nonhomogeneous linear ordinary differential equations. It

should be noted, however, that the solutions are not valid when $\omega_2 \rightarrow \omega_1$. This case requires a modification of the form of the solution assumed here for $x(t)$. The resolution of this situation remains a problem for further research.

REFERENCES

1. Krylov, N. and Bogoliubov, N. Introduction to Nonlinear Mechanics. Princeton: Princeton University Press (1947).
2. Bogoliubov, N. N. and Mitropolsky, Y. A. Asymptotic Methods in the Theory of Nonlinear Oscillations. New York: Gordon & Breach (1961).
3. Brunelle, E. J. The transient response of second order nonlinear equations. Int. J. Nonlinear Mechanics, 2, (1967) pp. 405-415.
4. Mendelson, K. S. Perturbation theory for damped nonlinear oscillations. J. Math. Phys., 11 (1970) pp. 3413-3415.
5. Anderson, G. L. An approximate analysis of nonlinear, nonconservative systems using orthogonal polynomials. J. Sound Vib., 29 (1973) pp. 463-474.
6. Anderson, G. L. Application of ultraspherical polynomials to nonlinear, nonconservative systems subjected to step function excitation. J. Sound Vib., 32 (1974) pp. 101-108.
7. Anderson, G. L. An asymptotic analysis of the response of nonlinear damped systems subjected to step function excitation. J. Sound Vib., 34 (1974) pp. 425-440.
8. Anderson, G. L. The method of averaging applied to a damped, nonlinear system under harmonic excitation. J. Sound Vib., 40 (1975) pp. 219-225.
9. Stanišić, M. M. and Euler, J. A. A contribution to the Krylov-Bogoliubov theory in nonlinear mechanics. Ing.-Arch., 42 (1973) pp. 371-380.

10. Kane, T. R. Free and forced oscillations of a class of mechanical systems. *Int. J. Nonlinear Mechanics*, 1 (1966) pp. 157-167.
11. Nayfeh, A. H. *Introduction to Perturbation Techniques*. New York: John Wiley (1981).
12. Sneddon, I. N. *Special Functions of Mathematical Physics and Chemistry*. Edinburgh: Oliver and Boyd (1961).

$\omega_1 = 1.5$, $\omega_2 = 6$, $b = 0.05$, $Q_0 = 1$, $\epsilon = 0.1$
 10-OCT-84 12:35:08

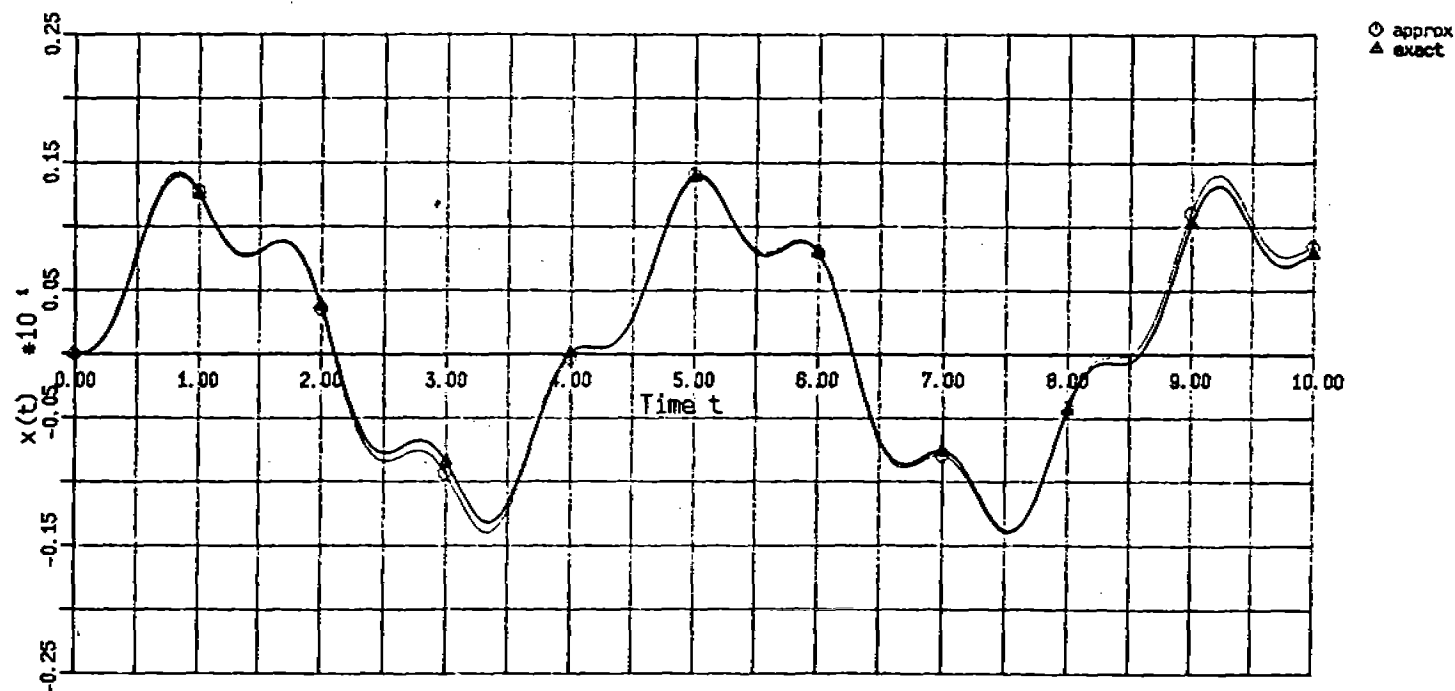


Figure 1. Variation of $x(t)$ versus t for $\omega_1 = 1.5$, $\omega_2 = 6$, $b = 0.05$, $Q_0 = 1$, and $\epsilon = 0.1$ according to (48).
 o-approximate solution, Δ -exact solution.

$\omega_1 = 1.5$, $\omega_2 = 4$, $b = 0.05$, $Q_0 = 1.0$, $\epsilon = 0.1$
 10-OCT-84 11:28:23

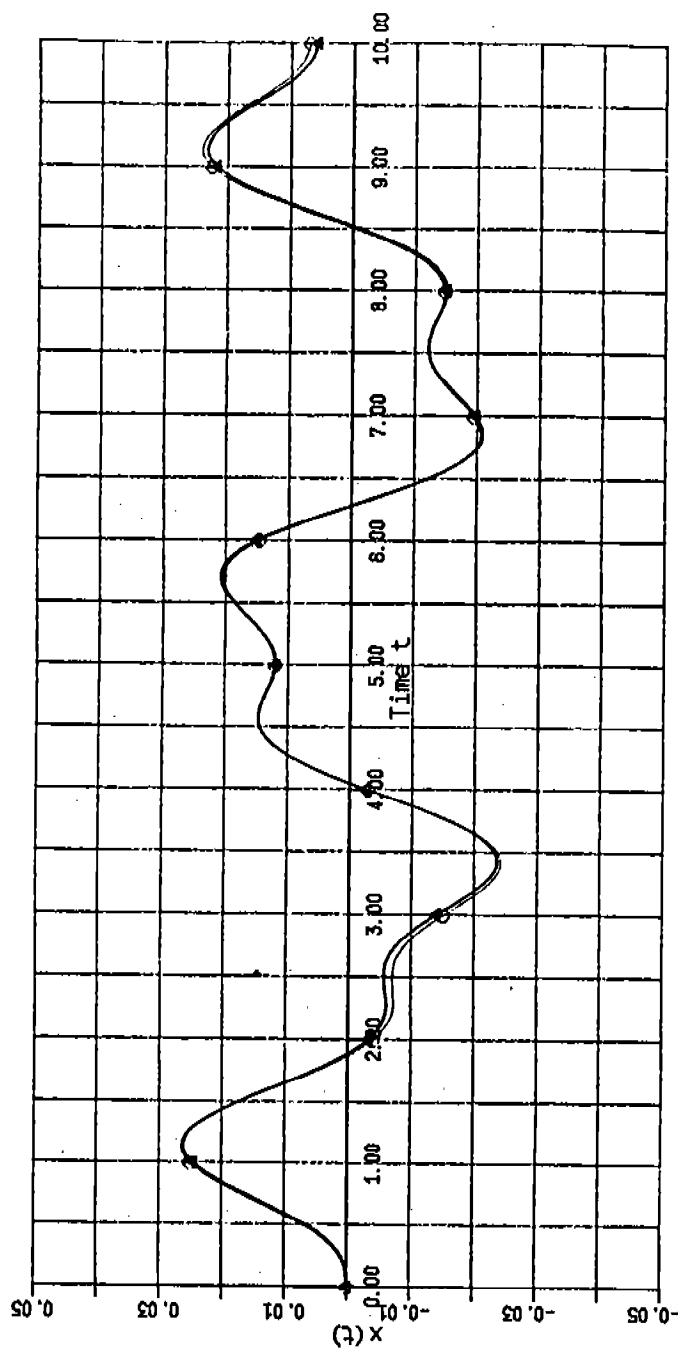


Figure 2. Variation of $x(t)$ versus t for $\omega_1 = 1.5$, $\omega_2 = 4$, $b = 0.05$, $Q_0 = 1$, and $\epsilon = 0.1$ according to (48).
 o-approximate solution, Δ -exact solution.

$w_1 = 1$, $w_2 = 6$, $\nu = 0.25$, $Q_0 = 1$, $\epsilon = 0.1$
 15-OCT-84 10:15:58

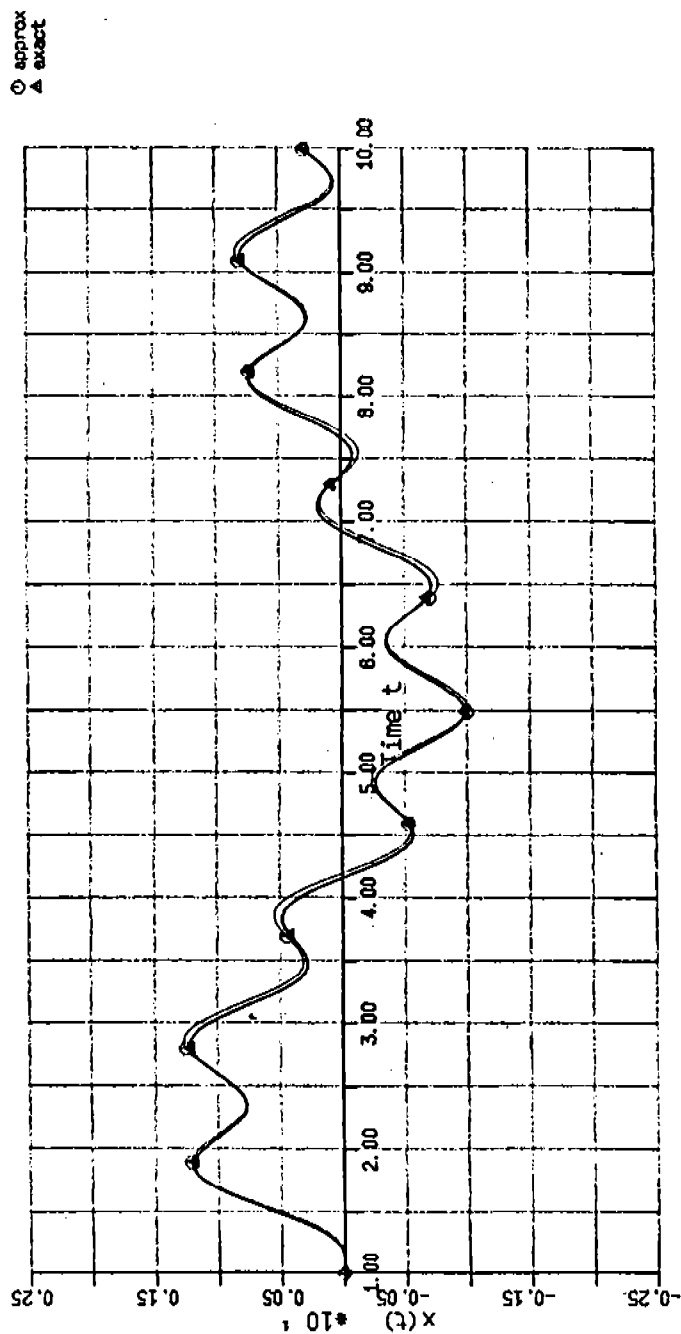


Figure 3. Variation of $x(t)$ versus t for $\omega_1 = 1$, $\omega_2 = 6$, $\nu = 1/4$, $Q_0 = 1$, and $\epsilon = 0.1$ according to (53).
 o-approximate solution, Δ -exact solution.

$w_1 = 1, w_2 = 4, \nu = 0.25, Q_0 = 1, \epsilon = 0.1$
 15-OCT-84 09:41:41

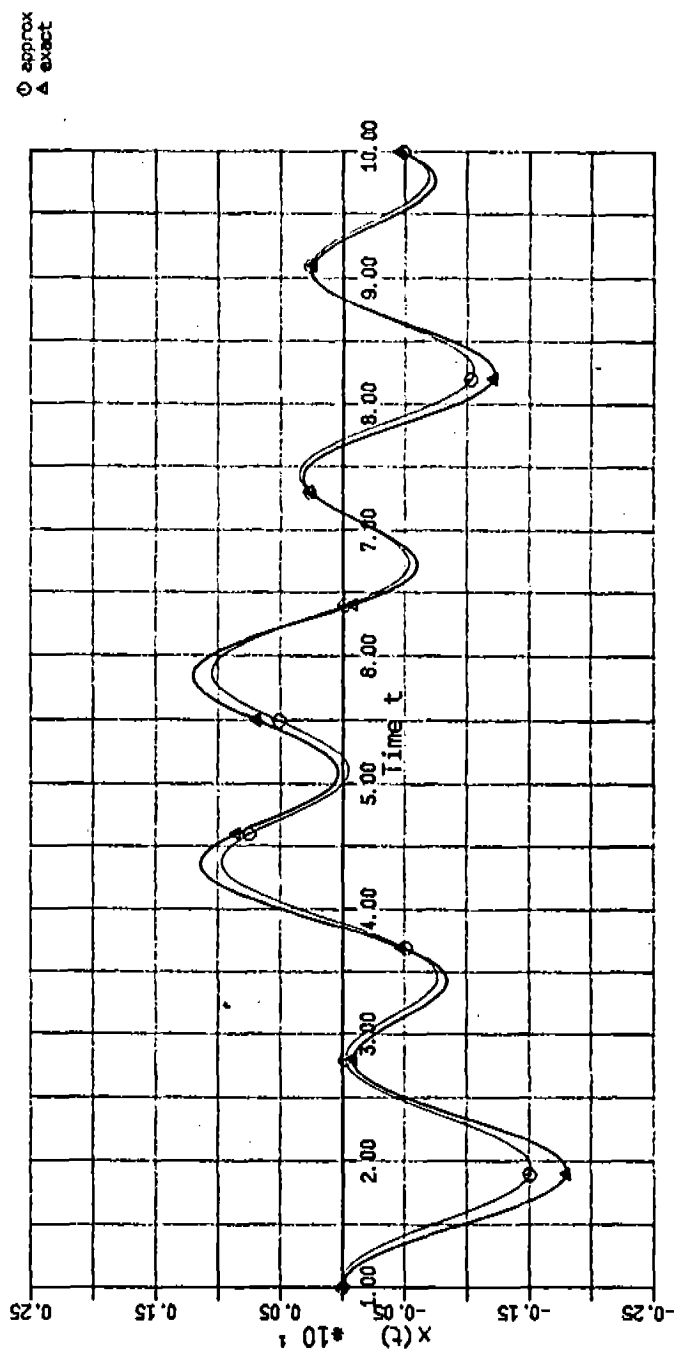


Figure 4: Variation of $x(t)$ versus t for $\omega_1 = 1, \omega_2 = 4, \nu = 1/4, Q_0 = 1$, and $\epsilon = 0.1$ according to (53).
 O-approximate solution, A-exact solution.

$\nu_2 = 6$, $\nu_1 = 3$, $\rho_0 = 1$, $\epsilon = 0.1$
 16-OCT-84 11:48:43

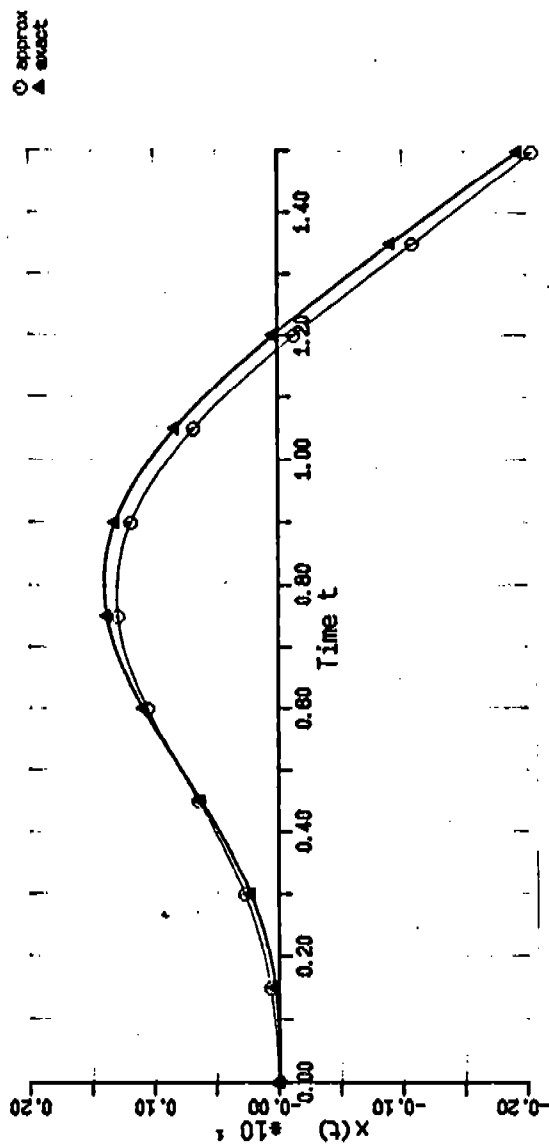


Figure 5. Variation of $x(t)$ versus t for $\omega_2 = 6$, $\nu = 3$, $Q_0 = 1$, and $\epsilon = 0.1$ according to (60).
 ○-approximate solution, ▲-exact solution.

$\omega_2 = 10$, $\nu = 6$, $Q_0 = 1$, $\epsilon = 0.1$
 16-OCT-84 13:18:06

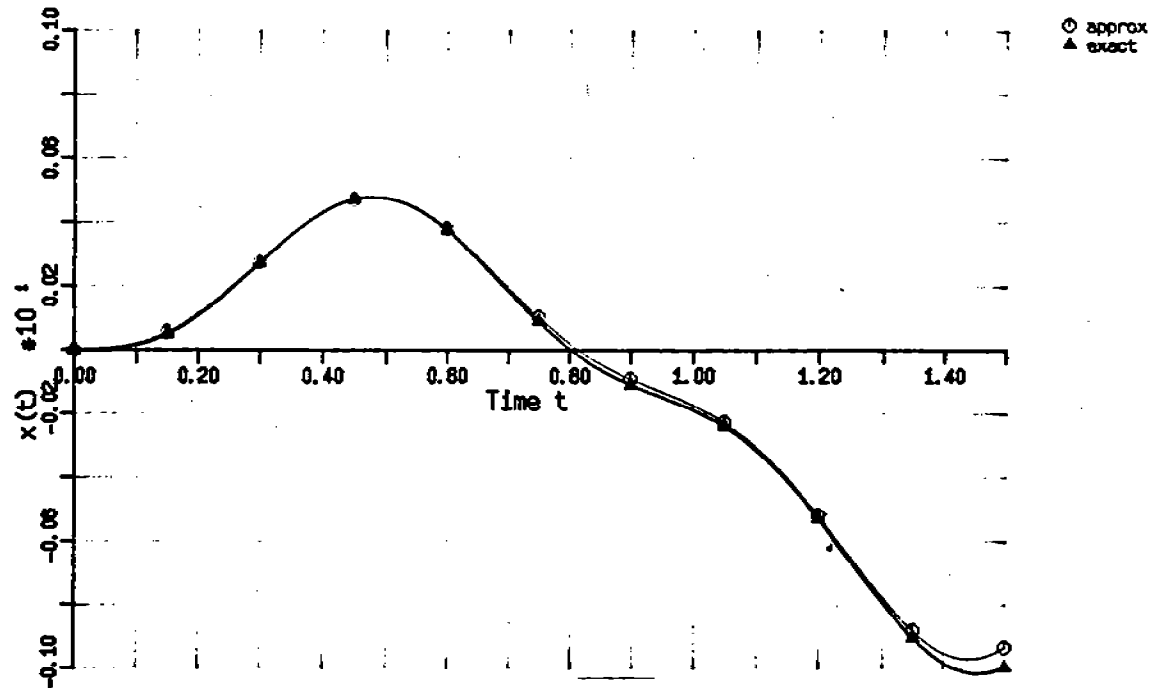


Figure 6. Variation of $x(t)$ versus t for $\omega_2 = 10$, $\nu = 6$, $Q_0 = 1$, and $\epsilon = 0.1$ according to (60).
 o—approximate solution, Δ —exact solution.

$\nu_1 = 1, \nu_2 = 6, Q_0 = 1, \epsilon = 0.1, g = 0.25, \alpha = 1$
 29-OCT-84 16:56:03

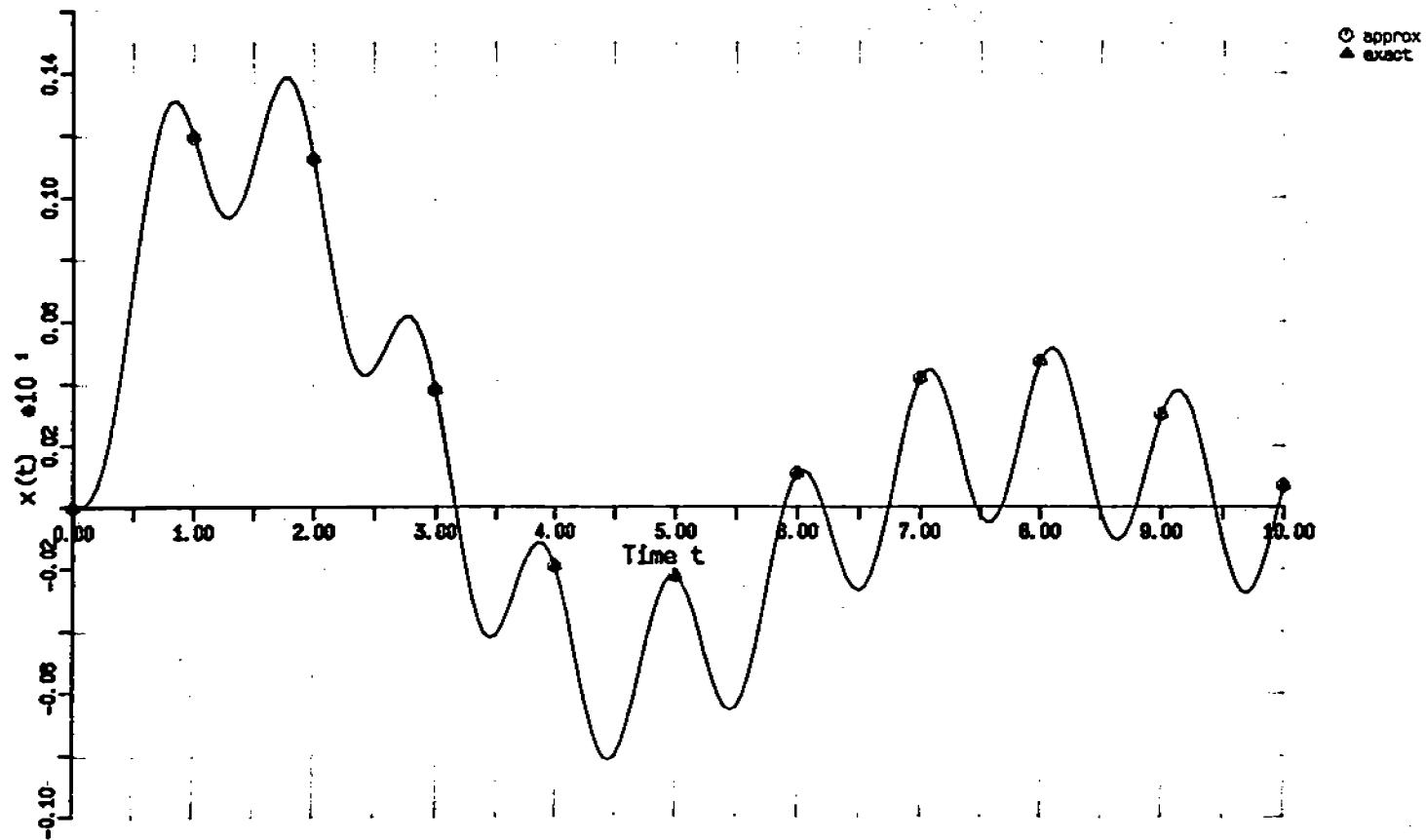


Figure 7. Variation of $x(t)$ versus t according to (88) for Duffing's equation with $\omega_1 = 1, \omega_2 = 6, Q_0 = 1, \epsilon = 0.1, g = 0.25$, and $\alpha = 1$. \circ -approximate solution, Δ -exact solution.

$v1 = 1, v2 = 6, Q_0 = 1, \epsilon = 0.1, g = 0.75, \alpha = 1$
 30-OCT-84 08:19:49

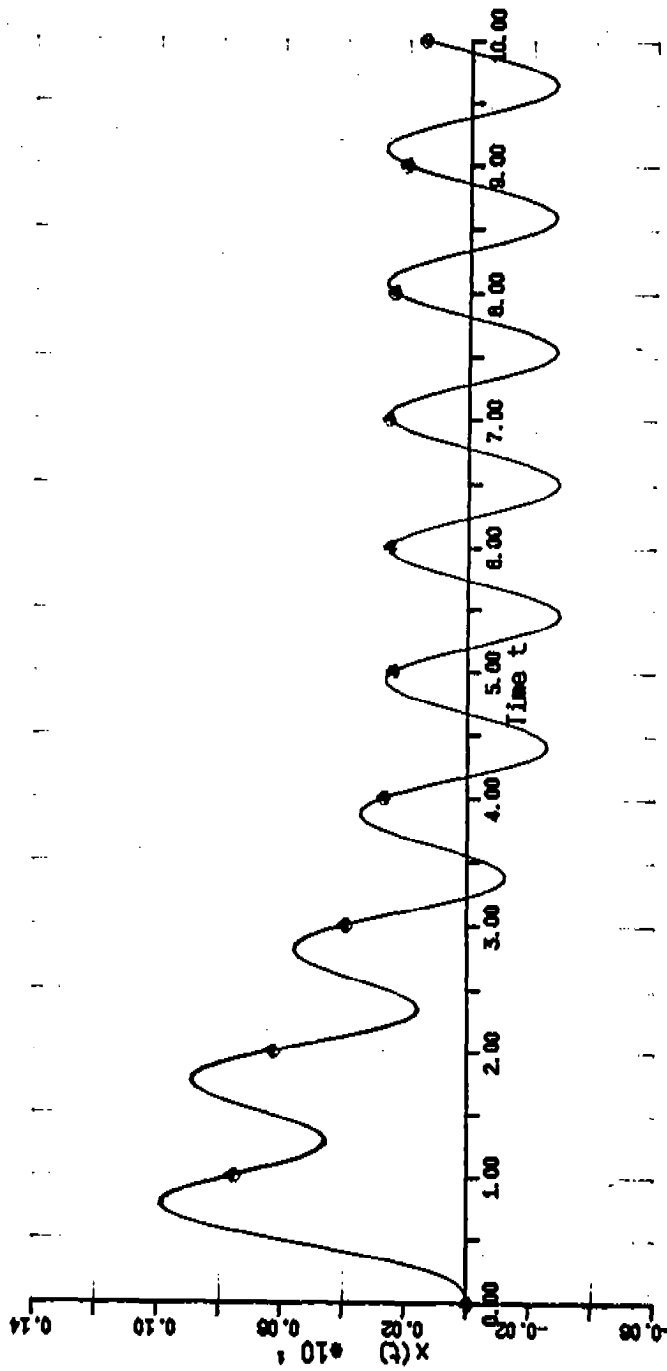


Figure 8. Variation of $x(t)$ versus t according to (88)
 for Duffing's equation with $\omega_1 = 1, \omega_2 = 6, Q_0 = 1, \epsilon = 0.1, \gamma_1 = 0.75$, and $\alpha = 1$. o-approximate
 solution, Δ -exact solution.

$v_1 = 1, v_2 = 4, Q_0 = 1, \epsilon = 0.1, g = 0.25, \alpha = -1$
 31-OCT-84 11:05:12

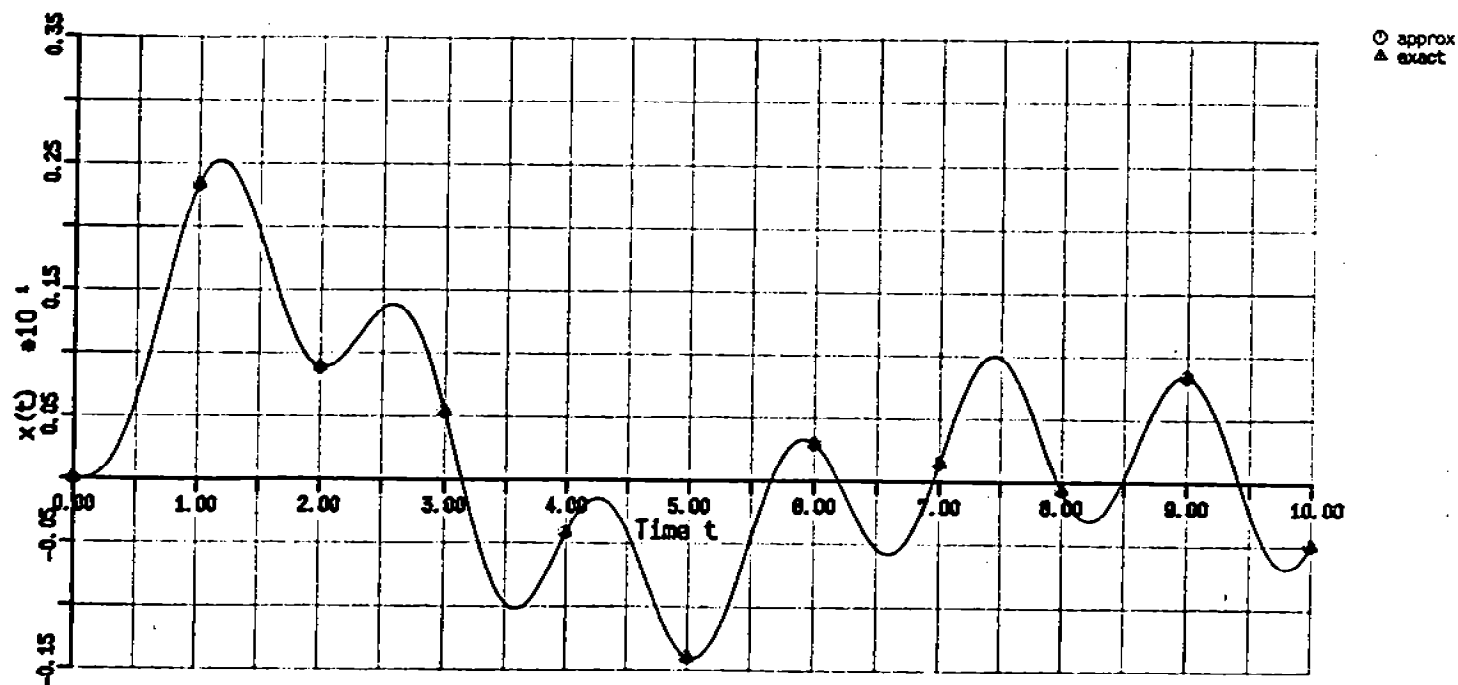


Figure 9. Variation of $x(t)$ versus t according to (99) for Duffing's equation with $\omega_1 = 1, \omega_2 = 4, Q_0 = 1, \epsilon = 0.1, \gamma = 0.25$, and $\alpha = -1$. \circ -approximate solution; Δ -exact solution.

$\omega_1 = 1, \omega_2 = 4, Q_0 = 1, \epsilon = 0.1, g = 0.75, \alpha = -1$
 31-OCT-84 10:46:47

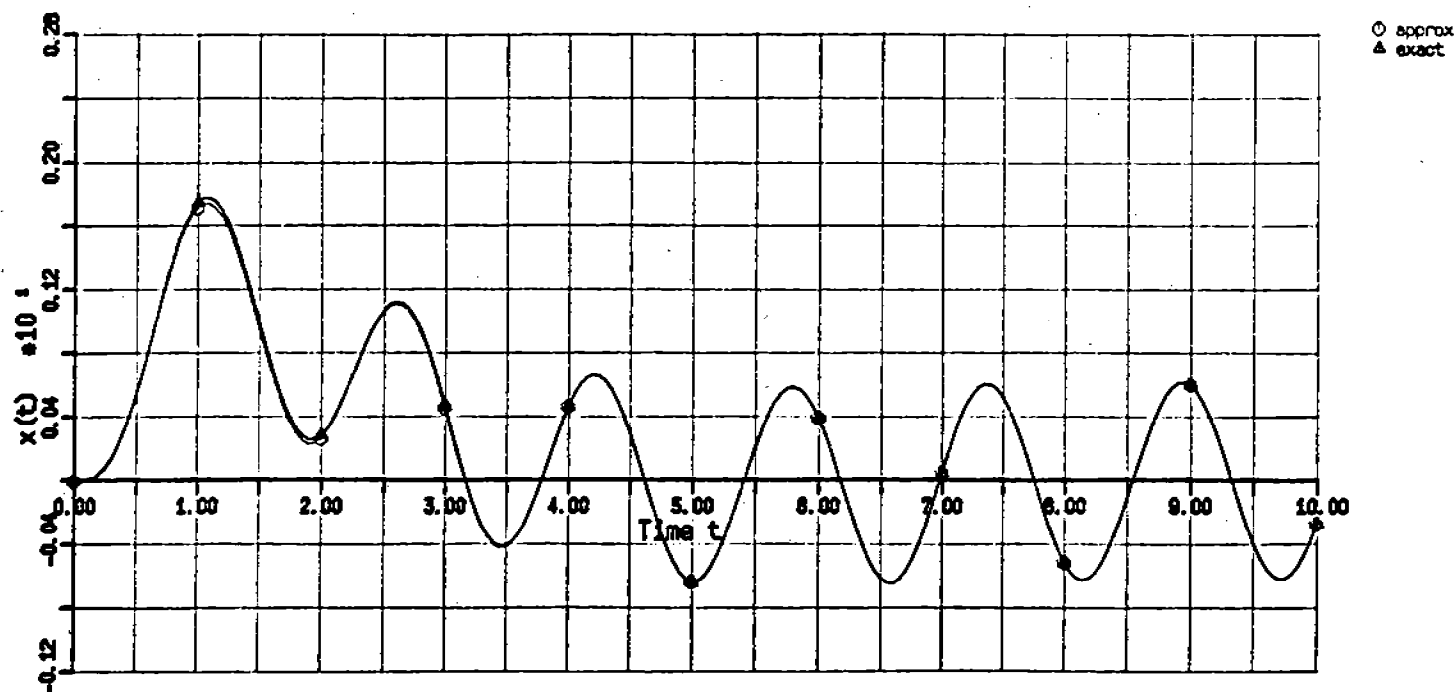


Figure 16. Variation of $x(t)$ versus t according to (99) for Duffing's equation with $\omega_1 = 1, \omega_2 = 4, Q_0 = 1, \epsilon = 0.1, \gamma_1 = 0.75$, and $\alpha = -1$. o-approximate solution, Δ -exact solution.

COMPUTATION OF RESIDUAL STRESSES DUE TO PHASE TRANSFORMATIONS
DURING QUENCHING OF HOLLOW CYLINDERS

J. D. Vasilakis

U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189-5000

ABSTRACT. In a previous paper, a method for computing the stresses due to the combined effects of transient temperatures and material phase transformations was described. The general purpose finite element program ADINAT/ADINA was used for the computation of both the transient temperatures and the associated stresses. The problem considered was that of an axisymmetric hollow cylinder undergoing a water-spray quench. The present work considers a similar model, but is better able to describe the residual stress state because of the availability of a more accurate set of properties for the material expansion due to the phase transformation. Effects on the transient and residual stresses due to modifications of the material expansion and varying quench rates are discussed.

It is found that the stresses due to the transformation are more severe than those due to the transient temperatures alone. Inelastic behavior is found to occur in all the cases considered and high residual stresses can exist on the inner and outer surfaces. While dependent on actual material composition, these residual stresses can lead to quench cracking.

The model describes the rapid quenching of steel gun tubes for the purpose of developing a martensitic grain structure and desired physical properties in the tube.

I. INTRODUCTION. Rapid quenching of components from initially high temperatures usually results in development of residual stresses within the components upon cooling. The quenching is normally undertaken to develop a desired microstructure in the material which would determine its behavior or response in use. These changes in microstructure, or transformations, cause volume changes which can give rise to stresses in the component in addition to the stresses due to the rapid temperature changes. Thus, the residual state of stress which exists when the component is cooled to room temperature is due to the combined effect of transient temperatures and material phase transformation.

In a previous paper [1], a method for computing the stresses due to these combined effects was described. The general purpose finite element program ADINAT/ADINA was used for the computation of both the transient temperatures and the associated stresses. The problem considered was that of an axisymmetric hollow cylinder undergoing a water-spray quench. The present work considers a similar model but is better able to describe the residual stress state because of the availability of a more accurate set of properties for the material expansion due to the phase transformation. The material

properties and system parameters used in the computations, such as quenching time, were chosen using the rotary forge quench facility at Watervliet Arsenal as a model.

The quench facility has nozzles on several diametral planes for spraying water on the outer diameter of a long tube as it is slowly rotated. The bore, or inner surface of the tube, can also be cooled by a bore flush. There have been several types of quench cycles in the past. These include varying quench times of the outer diameter for the breech and muzzle ends of the tube. The long tubes are usually of constant bore diameter and varying outer diameter with the larger being the breech end and the smaller the muzzle end. Quenching of the bore can be omitted, delayed, simultaneous with the outer diameter quench, etc. The goal was to develop the design properties in the tube without causing quench cracking due to the high residual tensile stresses at the bore.

Since the diameter of these tubes varies slowly, end effects are ignored and the tube is treated as a long axisymmetric cylinder. In the earlier work [1], the transient temperatures and combined stresses for the breech and muzzle ends were treated as two separate cases. This earlier work also considered the effects of the different quench cycles. The geometry used is that of the muzzle or smaller end of the long tube. It is also assumed that no bore quench takes place. This was done because most of the quenching currently being undertaken at the facility does not use the bore quench. Latent heat is ignored in the computation of the transient temperatures. This is not due to a limitation of the model, but to a lack of appropriate input.

Based on recent experimental work [2], the residual stresses can now be computed from realistic temperature-transformation curves. Also, because of information on the martensite transformation itself, the effect of new quench cycles on the cooling curves can be seen. The resultant residual stress distributions can then be discussed.

II. PROBLEM STATEMENT. Thermal and transformation stresses are computed for long hollow cylinders as they are being quenched. The effects due to the transient temperature distributions and the martensite transformation are both considered in the stress calculations. The thermo-physical properties are assumed to be temperature dependent. The residual stress distributions at the end of the quench cycles are presented. Both experimentally developed and assumed transformation-temperature curves are used, and the quench cycle is varied. A general purpose finite element program ADINAT/ADINA is used for the computations.

III. FINITE ELEMENT PROGRAM. The finite element geometry for the problem is shown in Figure 1, along with a simplified drawing of a gun tube. Eight node quadrilateral elements are used in the model. The present work shows results only for the muzzle end of the tube. In the earlier work [1], stress and temperature results for the breech end of the tube due to different quench cycles and an assumed transformation-temperature curve were presented.

The finite element program actually consists of two parts, one for computing temperatures, ADINAT, and one for computing stresses, ADINA. Each

program can stand alone, but when one wishes to compute thermal stresses using the same geometry, ADINAT produces a file which includes the temperatures at the node for each time step if the problem is a transient one. This ADINAT output file can then be used as input to ADINA for the stress computation.

In the program, the thermo-physical properties were considered as functions of temperature. The convection losses during the heat transfer portion of the computation are considered to be due to the temperature difference between the tube wall and ambient, which is assumed to be 18.3°C (65°F). For the computation of stresses, one has a choice of several material behavior models in ADINA. The one chosen for this work was Model 10 [3] which is applicable to the thermo-elastic-plastic solution of interest. The yield criterion assumed was the distortion energy criterion and the yield stress was assumed to be a function of temperature. No creep or hardening was assumed although the model allowed both to be incorporated.

To compute thermal stresses, the problem for the transient temperatures was solved using just ADINAT as indicated above. The special file created by ADINAT was then used as input to ADINA to compute the thermal stresses. However, in many cases in solving the temperature problem, time increments vary. Short-time increments are used during periods of large transients, and longer-time increments when the temperature gradients are not as severe. While ADINAT allows one to change time increments, ADINA does not. This difficulty was overcome by manipulating the temperature file used for stress computation so that with the restart capability, ADINA would see only one time increment during any one computation interval. Finally, the restart facility in ADINA [4] was altered so that a restart could be undertaken from any previous time instead of just the last completed step.

The computation of transformation stresses and combined thermal and transformation stresses can be treated like thermal stresses with little additional effort. The effect of the transformation, at least the aspect of it giving rise to stresses, is to create a volume change in the material. In this case the volume change is an increase, and it occurs when the temperature at a point in the material becomes equal to the martensite start (M_s) temperature and is completed when it reaches the martensite finish (M_f) temperature. If the transformation is assumed to be isotropic, then the linear expansion can be taken as one-third the volume change. Reference 1 describes what was done when the expansion due to the volume change was available as a separate quantity from the thermal expansion coefficient of the material. The current work takes advantage of the fact that the expansion is available from experiments as the combined effect.

IV. EXPERIMENTAL WORK. Cote [2] is conducting tests on various steels used in the manufacture of large caliber cannon. Small samples of the steels are quenched at different rates and the expansion in each sample is determined. He supplied the data for the experimental curve for the linear expansion versus temperature curve shown in Figure 2. He also suggested the modified curve as being typical of another type of steel or one in which some bainite has formed in addition to martensite. Other aspects of his work indicate that the desired microstructure can be achieved with slower quench cycles. This work has guided some of the decisions as to what quench cycles

to run, what transformation curves to use, etc.

V. RESULTS. The early work [1] was based on the assumed transformation-temperature curve shown in Figure 2. The experimental curve was not available at that time. Based on that curve, however, several different quench cycles were run to determine the transient temperature distributions and the associated residual stresses. The quench cycles previously run were:

1. Bore quench started at the same time as the OD quench.
2. Bore quench started 30 seconds before the OD quench began.
3. Bore quench started 30 seconds after the OD quench began.
4. No bore quench.

These were run for both the muzzle and the breech end of the gun tube. For comparison purposes, some of this earlier work is shown in Figures 3 through 5. The results are for the muzzle end with no bore quench taking place. Figure 3 shows the variation with time of the tangential stress at the bore and in the outer surface. Since this was a four minute quench simulation, the stresses at time 250 seconds are the residual stresses at those respective points. The breaks in the curves indicate when the phase transformation would begin and end in the bore and on the outer surface. The onset of inelastic deformation can also be found as can changes in the slope of the phase transformation-temperature curve. Figure 4 shows the residual stress distribution throughout the cylinder wall. High compressive stresses are seen at the bore and high tensile stresses on the outer surface. Figure 5 is a series of figures showing the tangential stress distribution across the wall thickness at different times. The transformation has started on the outer diameter by 150.5 seconds and continues into the cylinder. It begins on the ID by 175.5 seconds and the residual stress state (similar to Figure 4) is seen at time 245.5 seconds.

The results for other transformation curves in Figure 2 are shown in the remaining figures. The phase transformation curves used here for the computation of residual stresses are the experimentally determined one and the one modified from it. In addition, each was run with a fast quench (~ 4 minutes) and a slow quench (~ 15 minute) cycle. Boundary conditions for the computation of temperatures are not readily available. The only known information is that the temperature on the muzzle end of the tube reaches about 200°F in approximately four minutes. This point is about eight minutes on the breech end. A constant convection coefficient, h , is found by exercising the program which gives us this point and is used in subsequent runs. The geometry for the muzzle end was used and no bore quench was considered. Figures 6 through 9 represent the transient temperatures experienced by the tube during quenching. Figure 6 is the response of the fast quench and the temperature distribution throughout the wall at selected times is shown. The effect on the temperature distributed due to the OD only quench is easily seen, especially at early times. The difference between the outer diameter and bore diameter temperatures at any time is found in Figure 7, again for the fast quench cycle. Figures 8 and 9 show the thermal results for the slow quench cycle, with Figure 8 showing the radial distribution of temperatures, and Figure 9 the OD and ID temperature as they vary in time. It

is easily seen that the temperature difference between the ID and OD is much less during the slow quench.

Figures 10 and 11 show some of the stress results due to the experimental phase transformation curve and a fast quench cycle. This is perhaps the worst case as far as high stress gradients and residual stresses are concerned. The distribution of tangential stresses throughout the wall is shown in Figure 10 for specific times. At time 121 seconds the transformation has begun on the OD, and at time 150 seconds it has begun on the ID. The effect on the stresses is due primarily to the volume change associated with the transformation. This can easily be seen by considering the stresses prior to the transformation beginning and the changes to the state of stress after it has been completed. Figure 11 shows the fluctuation of the stresses at the ID and OD of the tube throughout this quench cycle. Again the breaks in the curve are due to the onset of transformations, the onset of inelastic material behavior, and changes in slope of the transformation curve. High tensile stresses are found at the bore. Figures 12 and 13 show the results for the experimental transformation curve and the slow quench. Since it was thought that the desired material microstructure would be developed even during a slow quench for a specific steel, the combination of loads was run. From Figure 12 one can see that the initial thermal stresses are small, as the thermal gradients are much less during the slower quench. The transformation on the OD started about 460-470 seconds into the run and on the ID probably about 40-50 seconds later. The residual state of stress is shown at the end of the quench cycle. High residual stresses are found at both the ID and OD but do not appear to penetrate into the interior of the tube section as they did during the rapid quench cycle. Bore tangential stress, however, is still tensile. Figure 14 shows the history of the tangential stress at the bore and OD. Most of the action occurs during the period from 450-625 seconds when the material is undergoing the transformation.

By subjecting the material to more time at the austenitizing temperature or perhaps by the changing the austenitizing temperature, it may be possible to change the shape of the phase transformation-temperature curve. For this reason, it was decided to look at the type of results that would arise if the modified transformation curve in Figure 2 was used. Figures 14 and 15 show the results for this curve with the first quench cycle. While the high stresses do occur on the ID and OD, the penetration of these stresses into the interior is not as much as it was with the experimentally determined transformation curve. Figures 16 and 17 depict the results for the modified transformation curve and slow quench. In this case, the residual stresses are small.

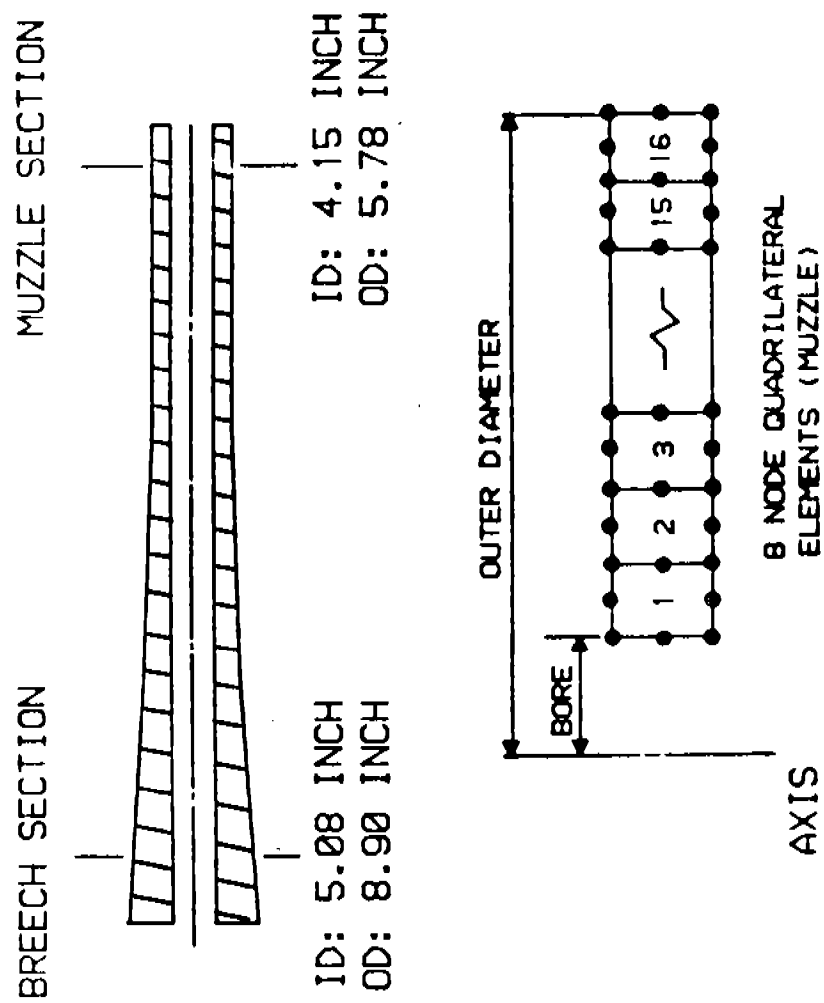
VI. CONCLUSIONS. Using experimental results as a guide, this work looked at the residual stress states in a long hollow tube which arise when the cylinder is quenched. These stresses are due to the transient temperatures during quenching and the material transformation that occurs. High tensile stresses occurred at the bore for all runs. These are the stresses that can lead to quench cracking.

In all runs, the stresses that occurred due to the phase change were much more severe than those due to the thermal gradients. During the slow quench, the thermal stresses were very small by comparison. For the more rapid quench, although more severe, they still were much less than the transformation stress. One should recall that the cylinder modeled is a steel one and the thermal conductivity is high tending to keep the gradients small. Varying the quench cycle, e.g., introducing bore quench, would not significantly change the stresses due to the temperature gradients alone.

The stresses due to the transformation cause inelastic material behavior, almost from the time the transformation begins. At the higher temperatures, it is expected that yielding could occur without generating cracks due to a more ductile response. A slow quench with the experimentally determined transformation curve or either quench with the modified transformation curve shows more shallow areas on the ID and OD for high residual stresses. Less material is thus subject to high stresses. Although this analysis cannot predict the onset of cracks due to quenching, it would tend to indicate that fast quench and experimental transformation curve would have a higher propensity for cracking. Either modifying the transformation curve or slowing the quench would help decrease the possibility of quench cracking.

REFERENCES

1. J. D. Vasilakis, "Thermal and Transformation Stresses in Hollow Tubes During the Quenching Process," Transaction of the First Army Conference on Applied Mathematics and Computing, ARO Report 84-1.
2. P. Cote, Benet Weapons Laboratory, Private Communication, September 1984.
3. M. Snyder and K.-J. Bathe, "Formulation and Numerical Solution of Thermo-Elastic-Plastic and Creep Problems," NTIS: PB-274-044, June 1977.
4. Fred Gregory, Ballistics Research Laboratory, Private Communication, April 1983.



PROBLEM GEOMETRY

FIGURE 1.

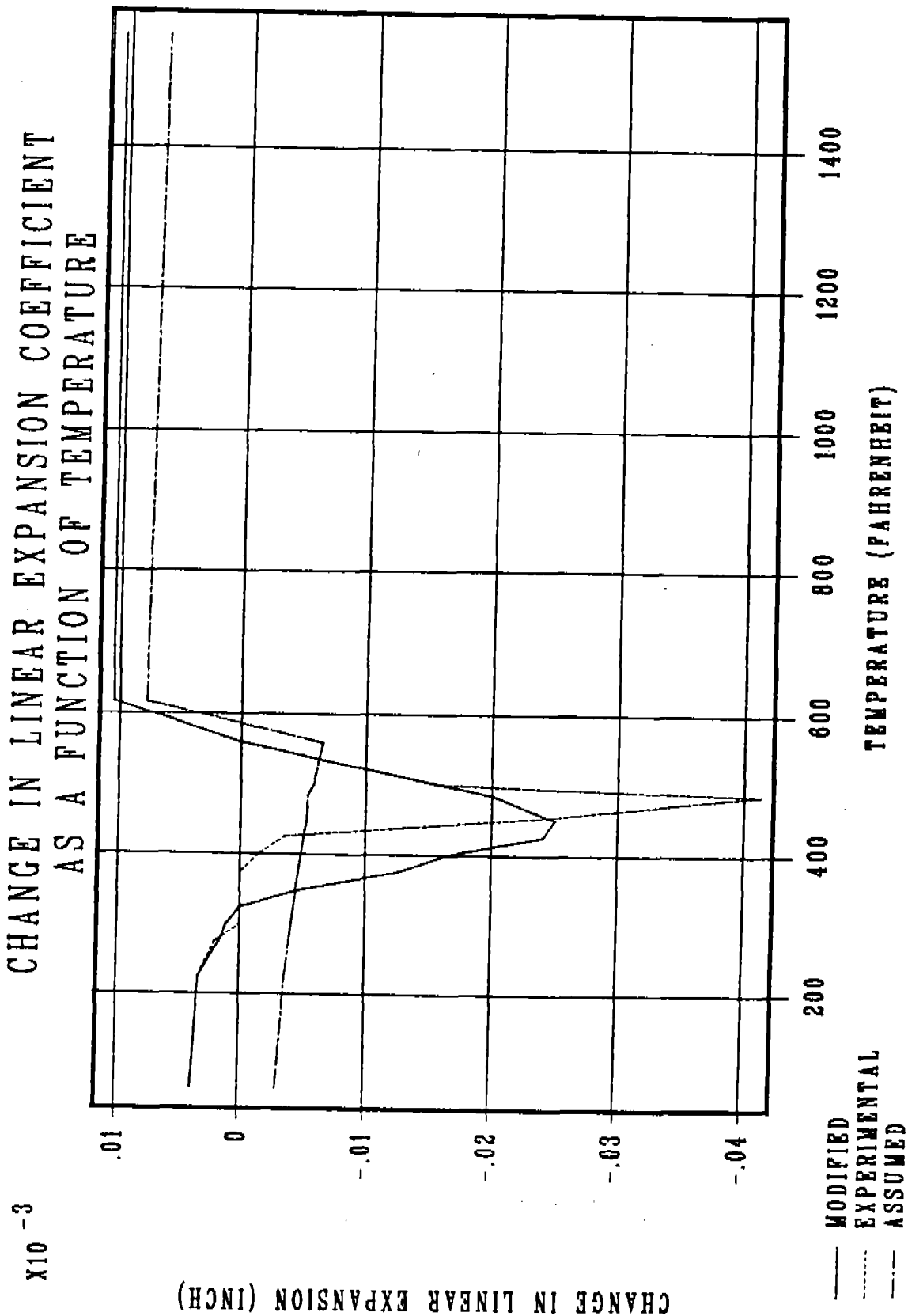


FIGURE 2.

(8/14/85)

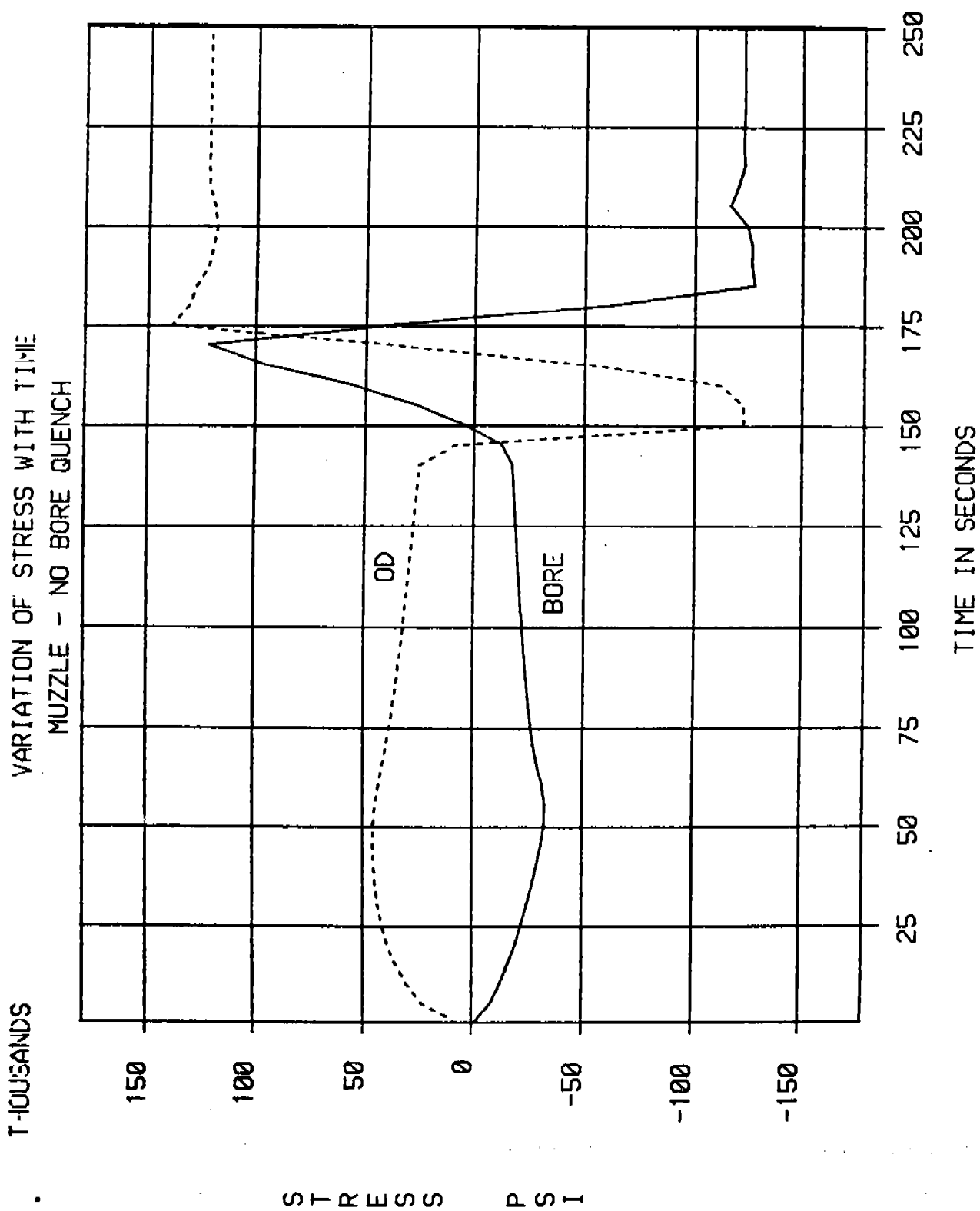


FIGURE 2

THERMAL AND TRANSFORMATION STRESS NO BORE QUENCH

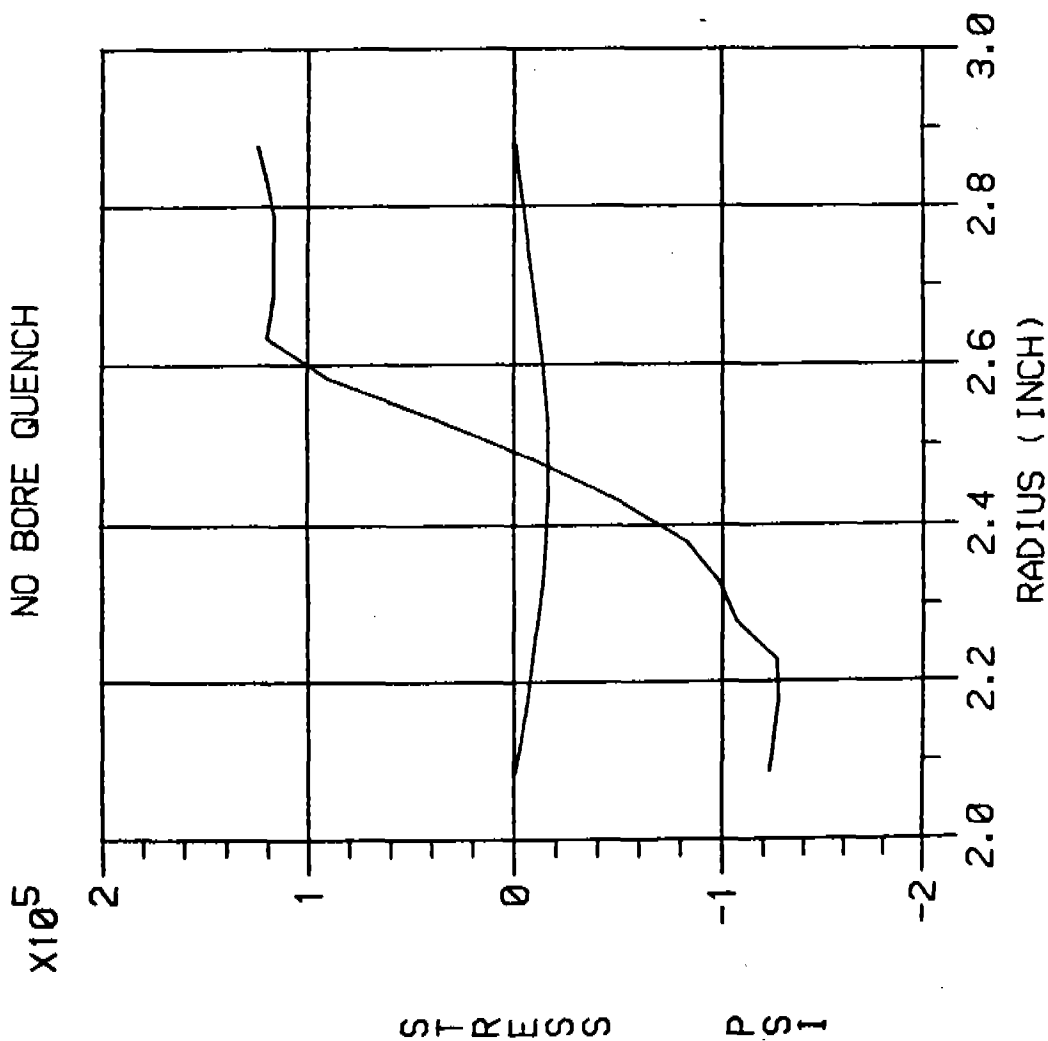


FIGURE 4.

TANGENTIAL STRESS vs RADIUS DURING QUENCH ASSUMED EXPANSION CURVE

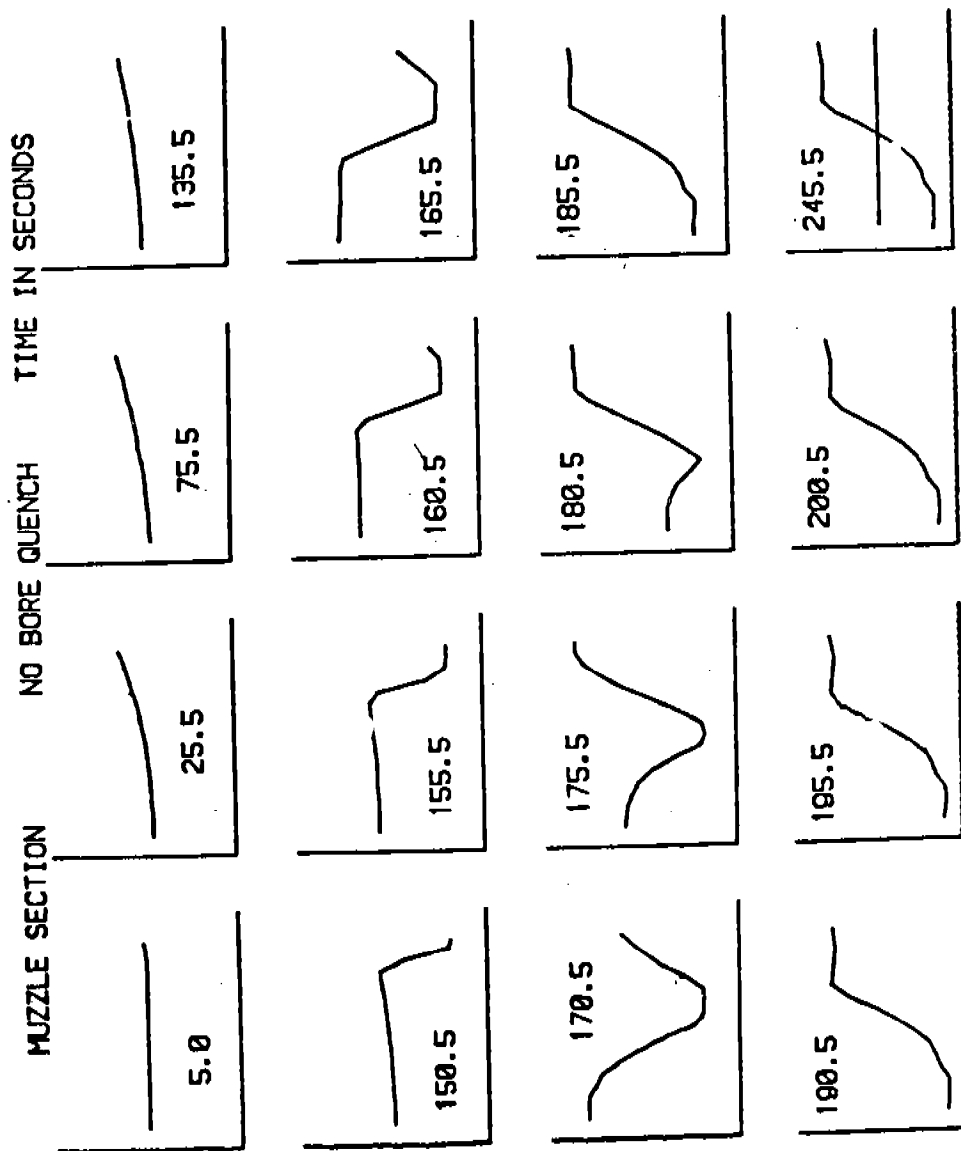


FIGURE 5.

TRANSIENT TEMPERATURES IN TUBE DURING QUENCH

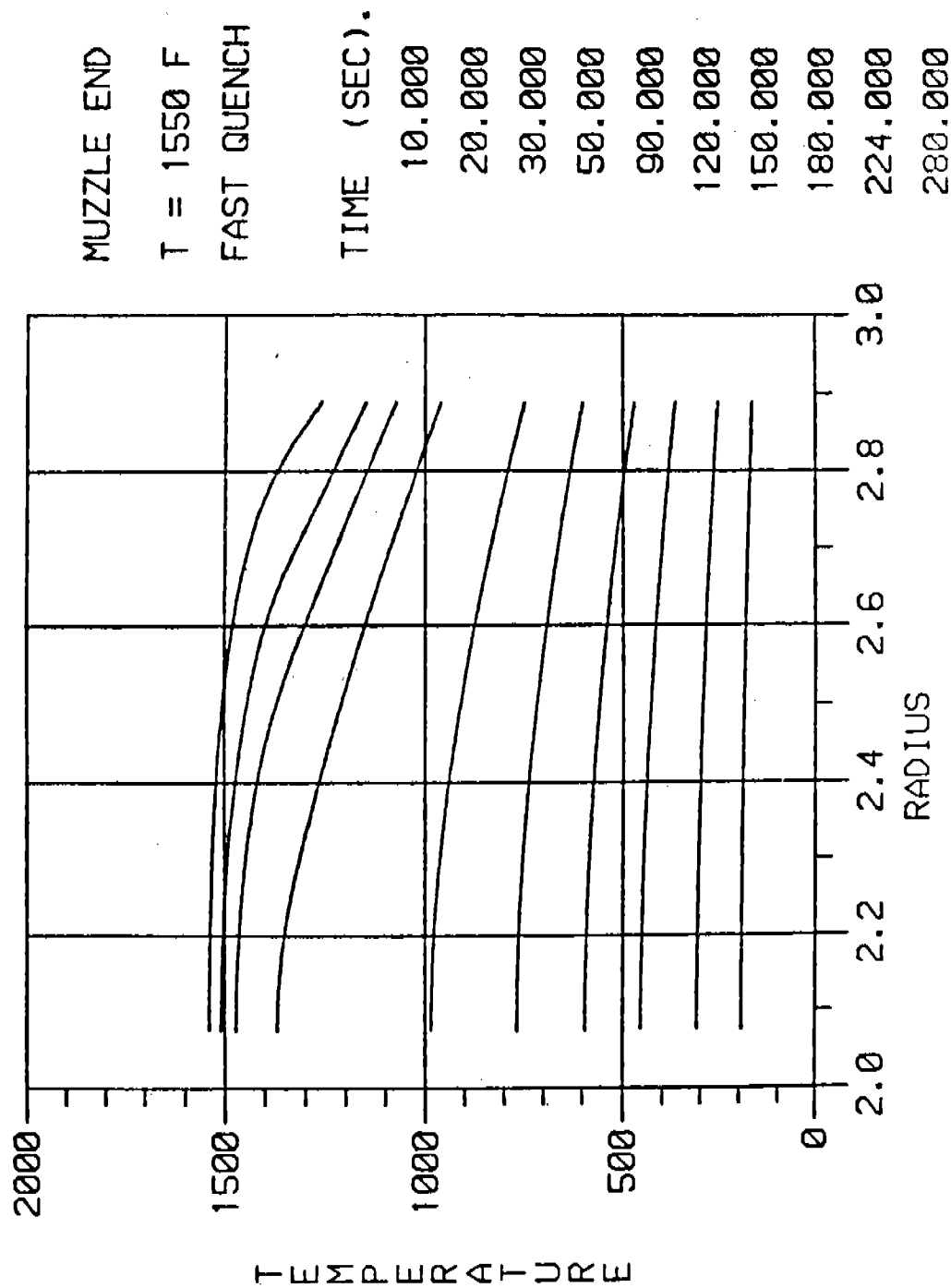


FIGURE 6.

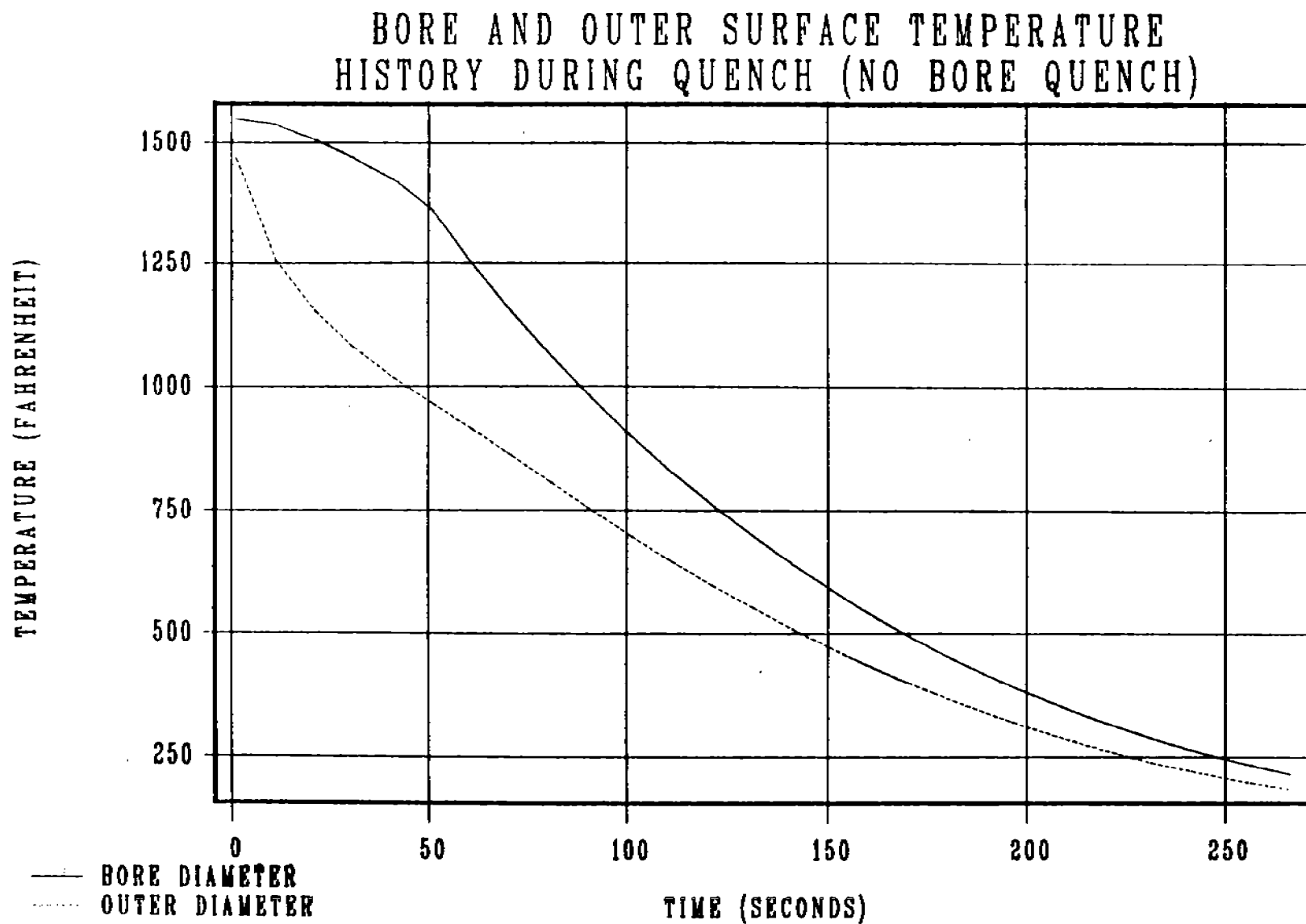


FIGURE 7.

TRANSIENT TEMPERATURES IN TUBE DURING QUENCH

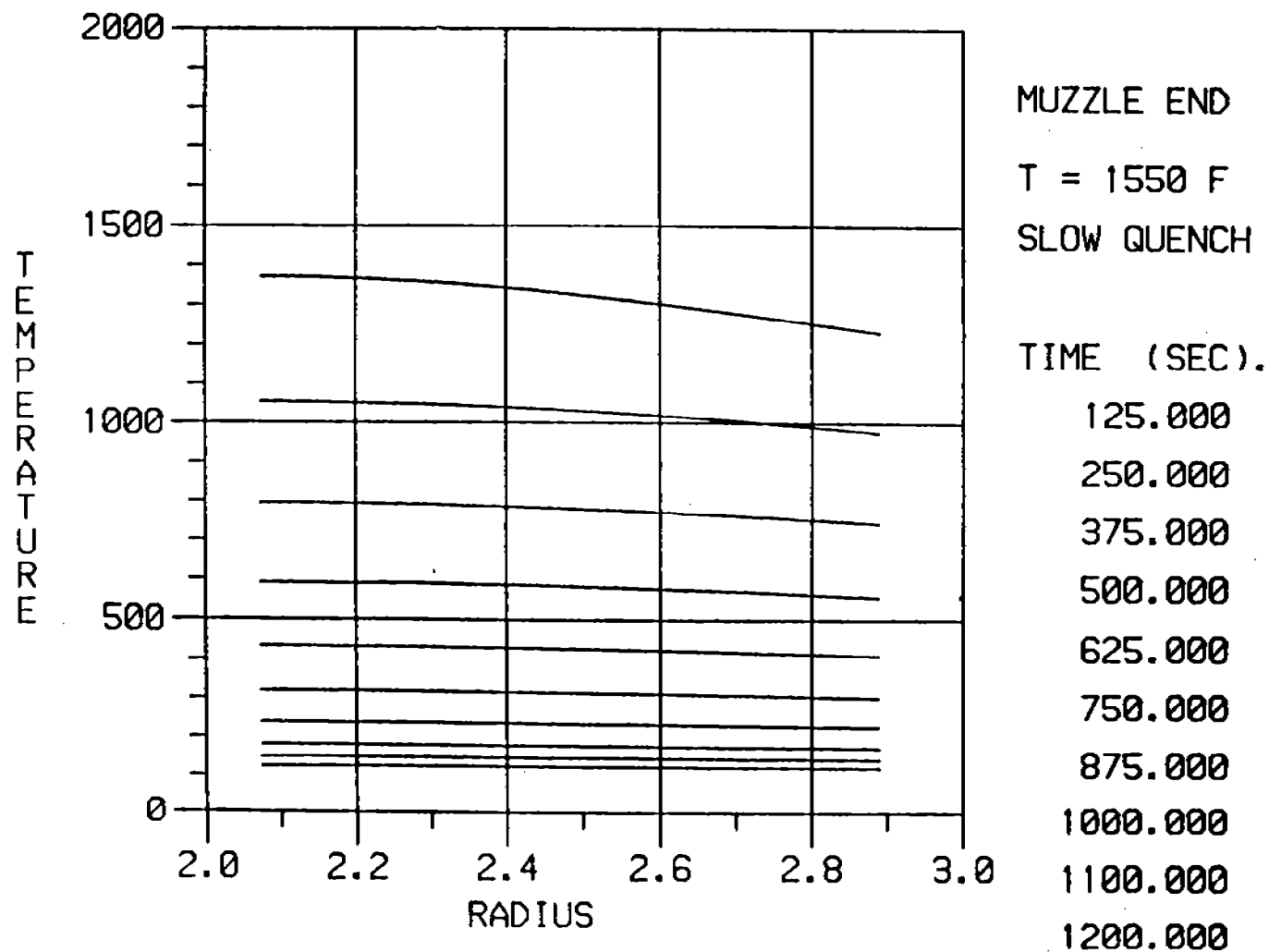


FIGURE 8.

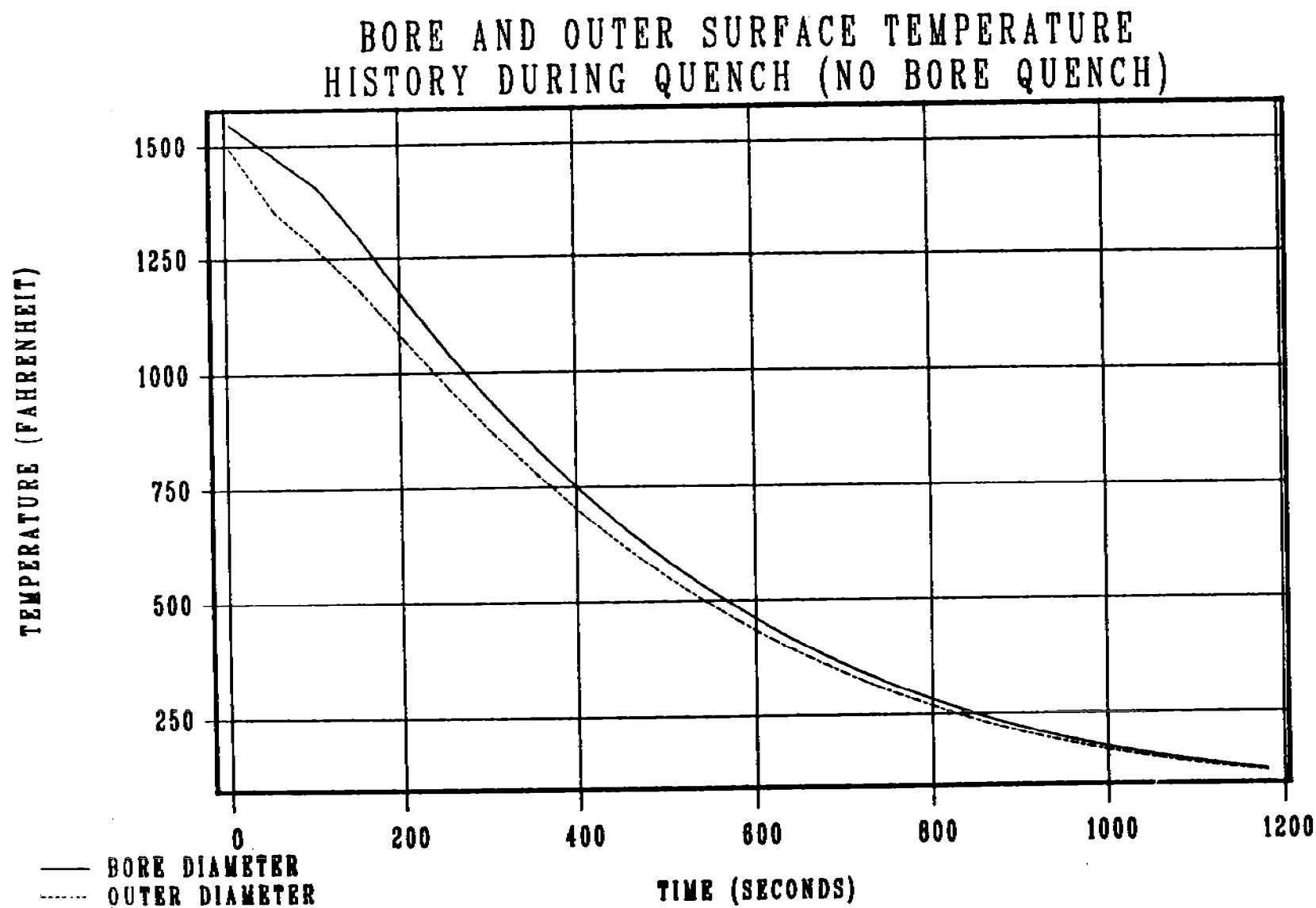
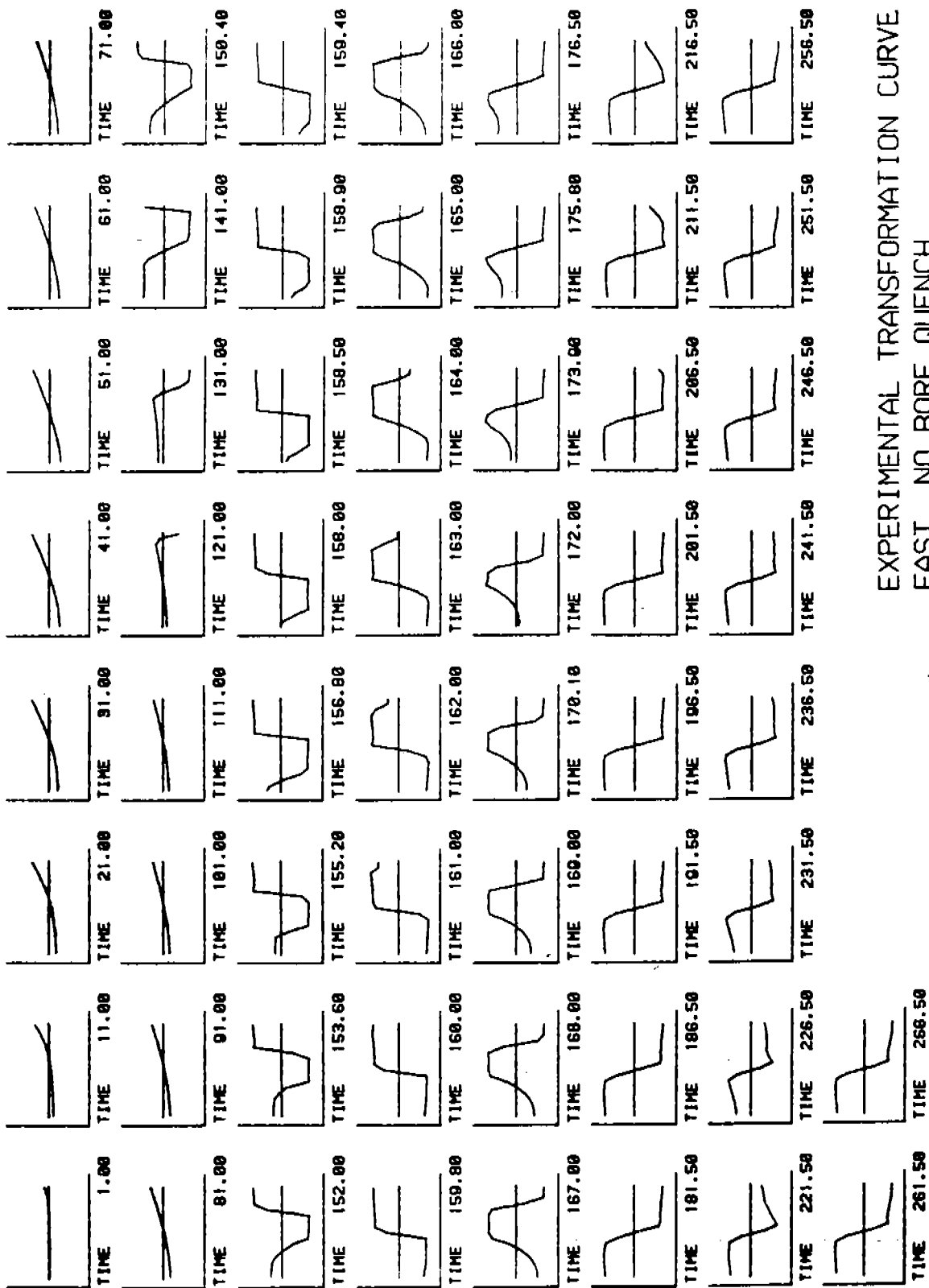


FIGURE 9.

SLOW QUENCH CYCLE
(10/5/84)

TANGENTIAL STRESS VS RADIUS FOR SPECIFIC TIMES DURING QUENCH



EXPERIMENTAL TRANSFORMATION CURVE
FAST, NO BORE QUENCH

FIGURE 10.

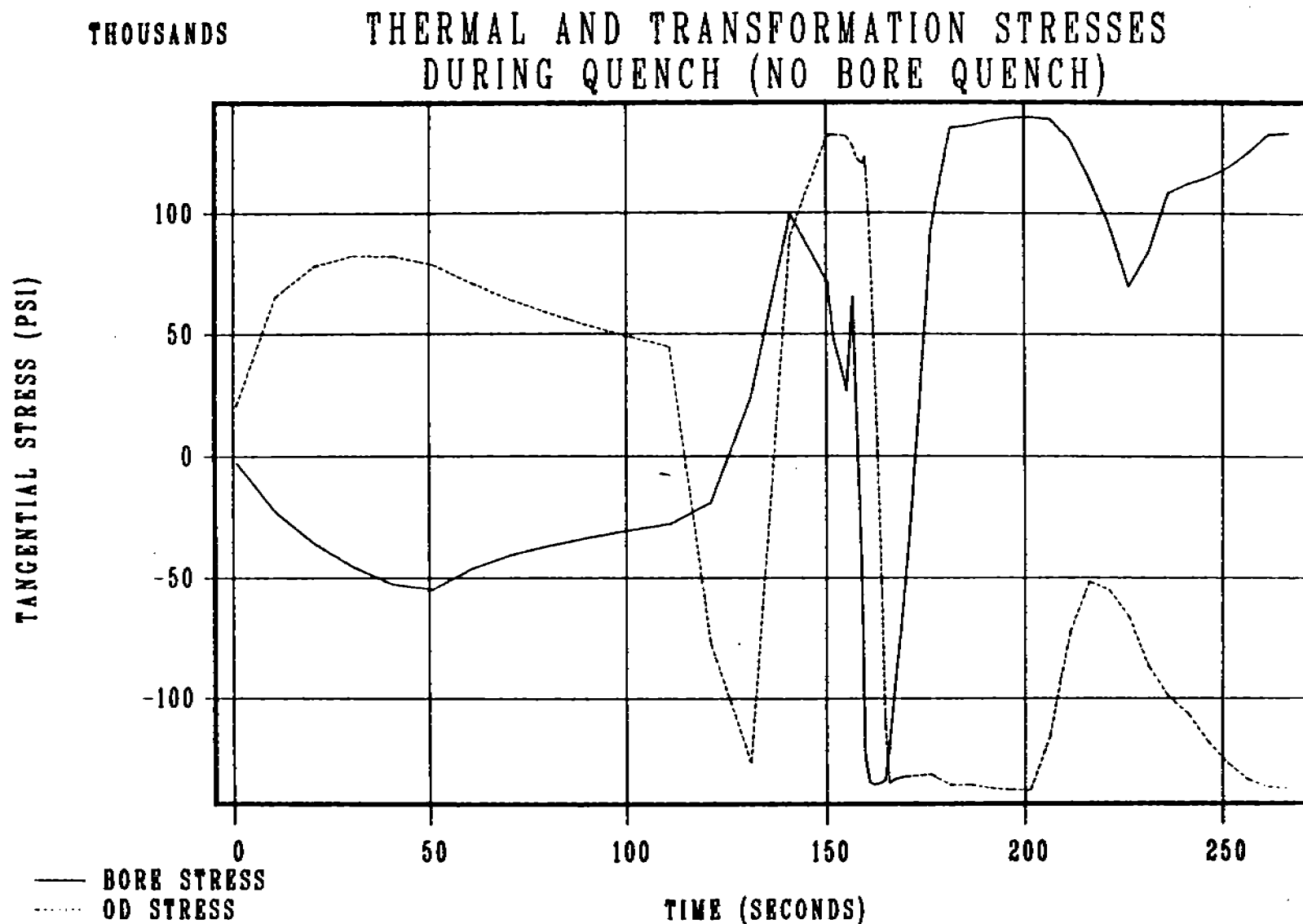
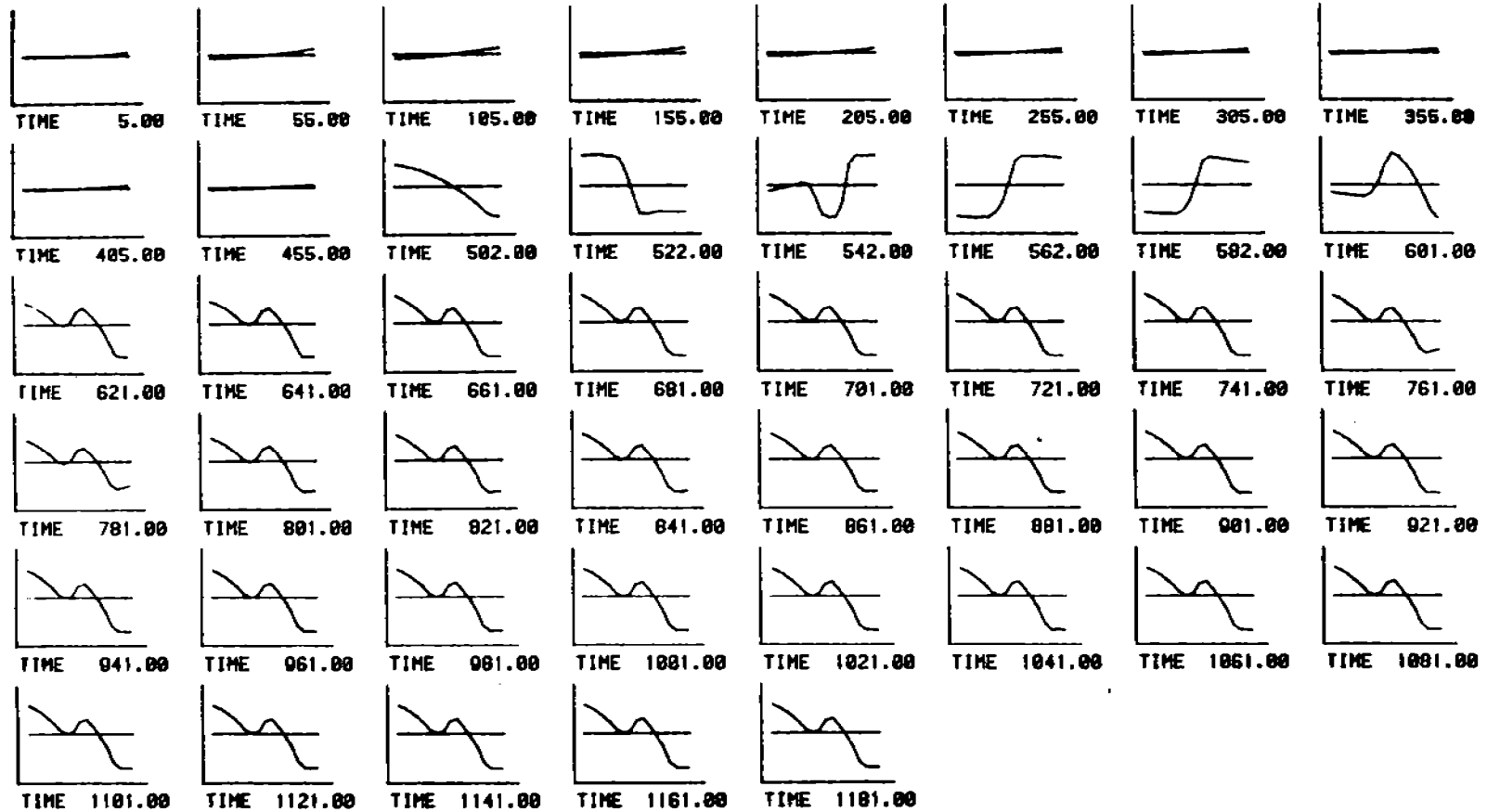


FIGURE 11.

BASED ON EXPERIMENTAL
TRANSFORMATION CURVE
(FAST QUENCH) 12/31/84

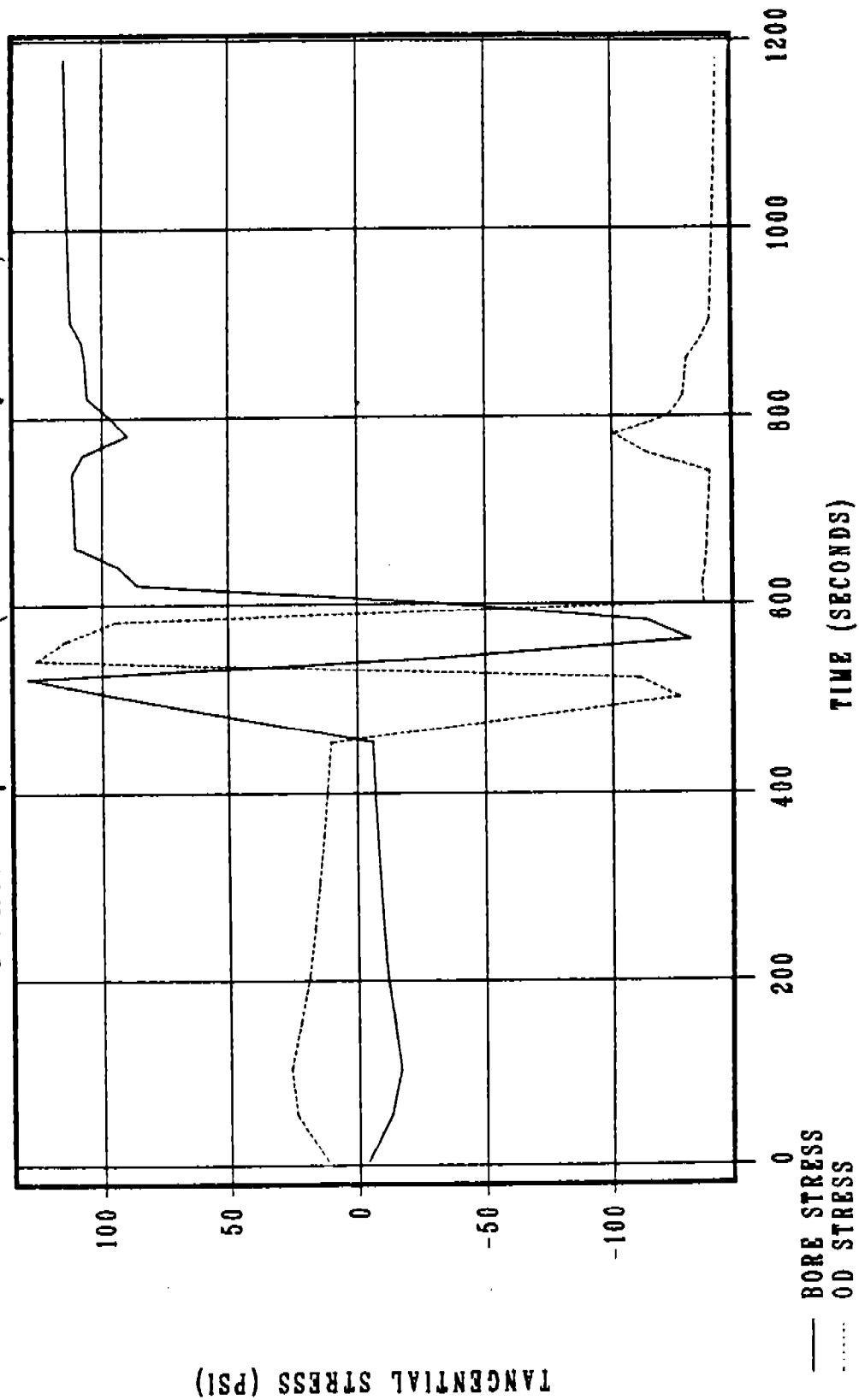
TANGENTIAL STRESS vs RADIUS FOR SPECIFIC TIMES DURING QUENCH



EXPERIMENTAL TRANSFORMATION CURVE
SLOW, NO BORE QUENCH

FIGURE 12.

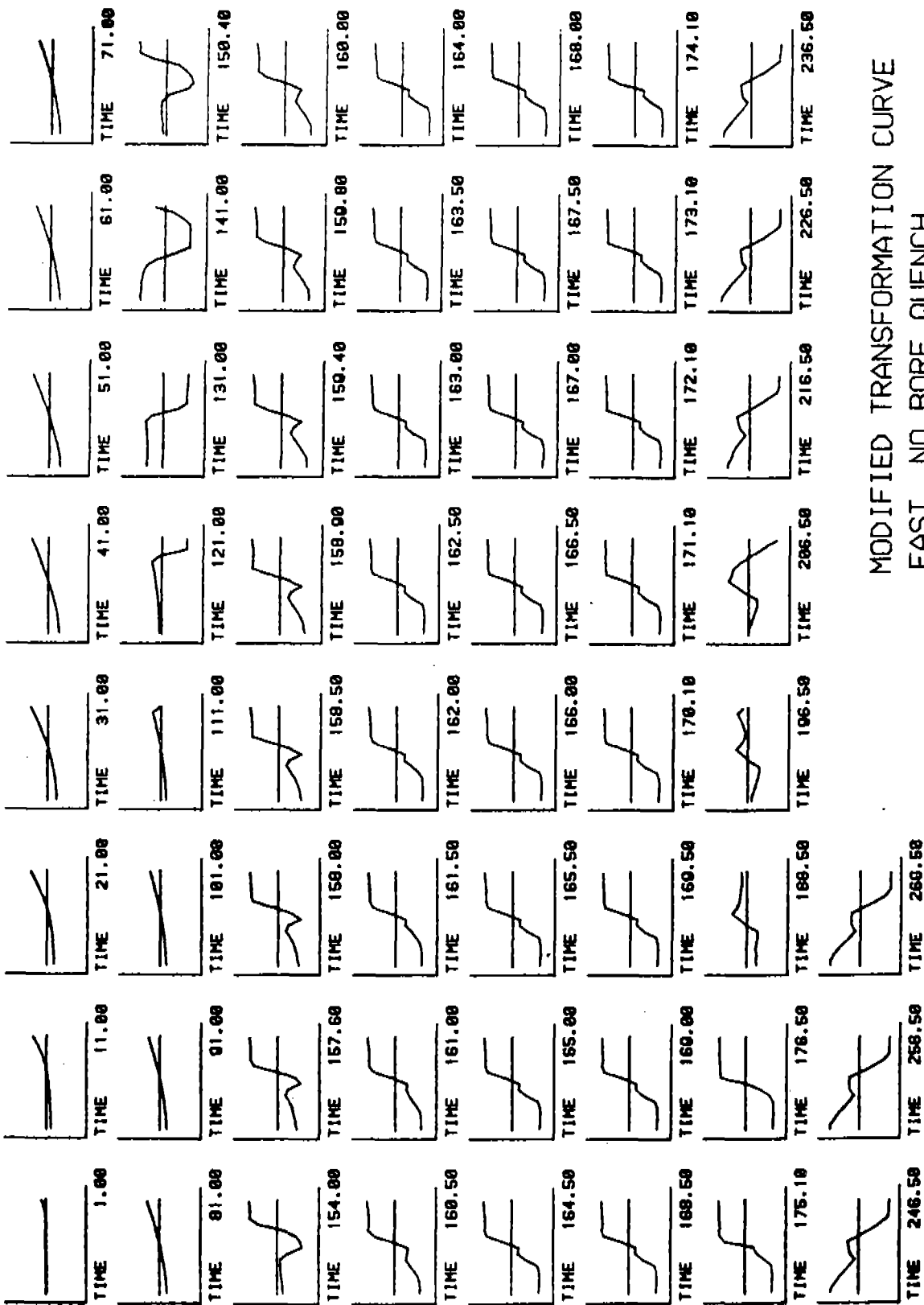
THERMAL AND TRANSFORMATION STRESSES DURING QUENCH (NO BORE QUENCH)



BASED ON EXPERIMENTAL
 TRANSFORMATION CURVE
 (SLOW QUENCH) 12/31/84

FIGURE 13.

TANGENTIAL STRESS VS RADIUS FOR SPECIFIC TIMES DURING QUENCH



MODIFIED TRANSFORMATION CURVE
FAST, NO BORE QUENCH

FIGURE 14.

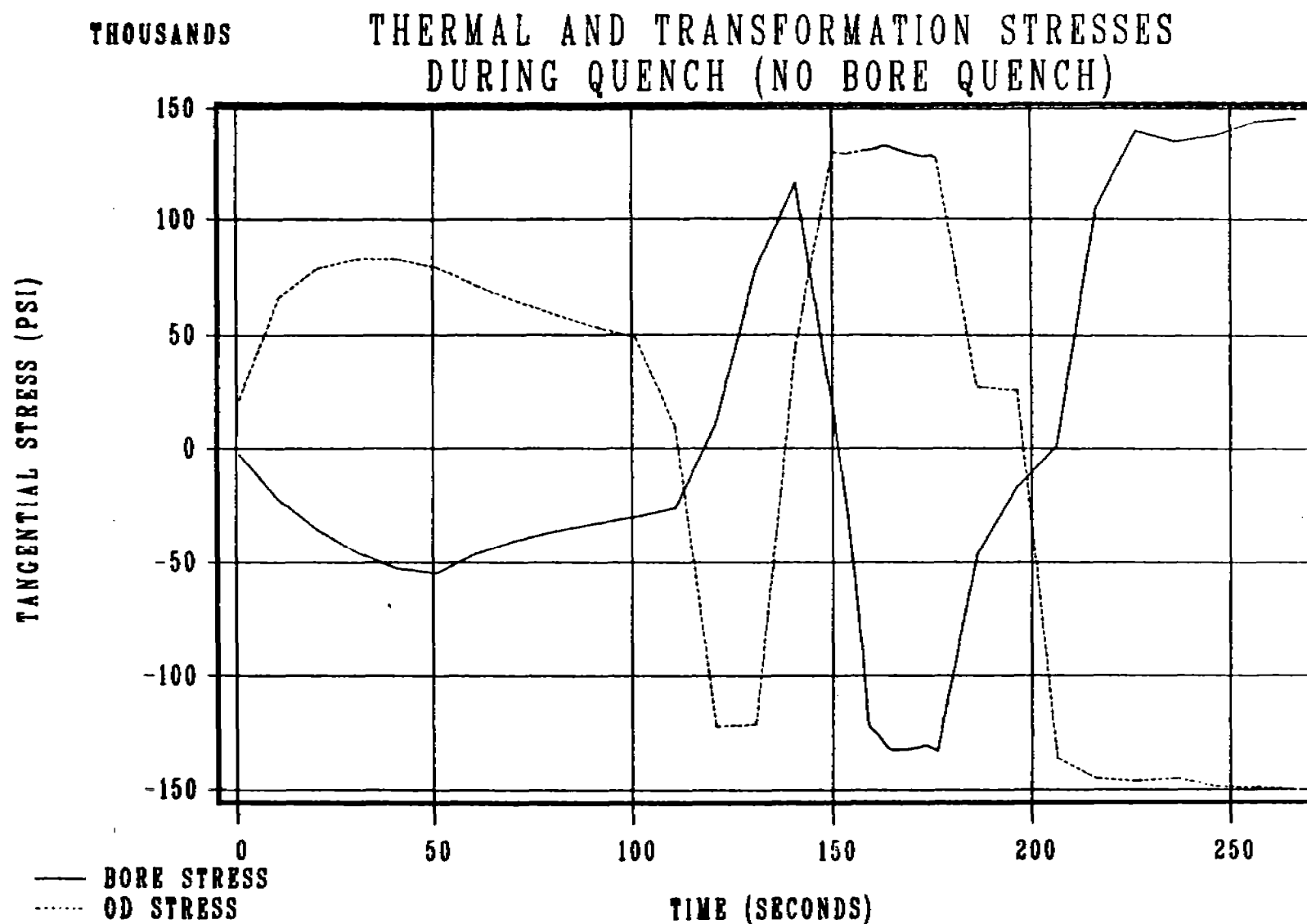
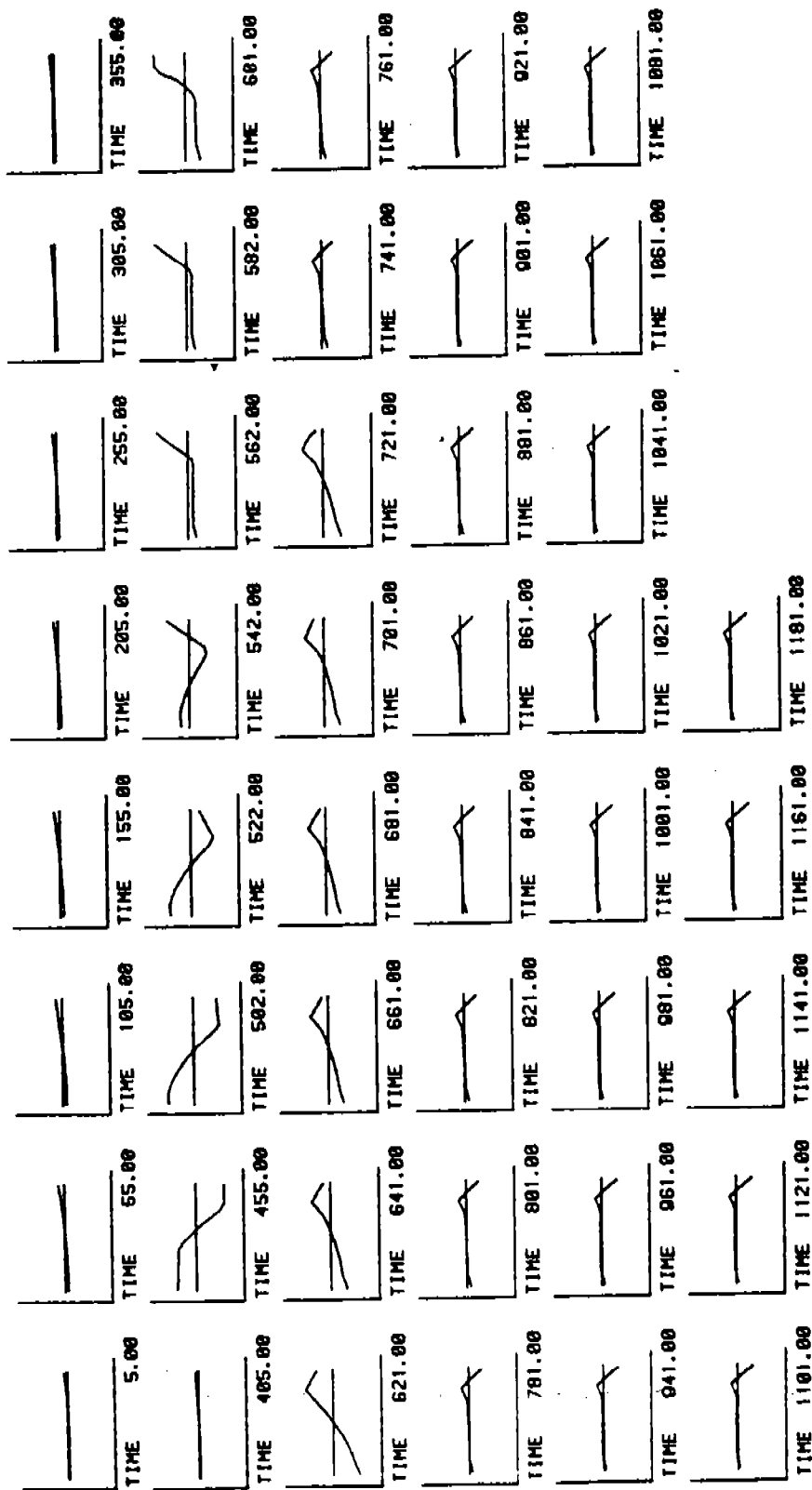


FIGURE 15.

BASED ON MODIFIED
TRANSFORMATION CURVE
(PAST QUENCH) 12/31/84

TANGENTIAL STRESS vs RADIUS FOR SPECIFIC TIMES DURING QUENCH



MODIFIED TRANSFORMATION DATA
SLOW, NO BORE QUENCH

FIGURE 16.

THOUSANDS

THERMAL AND TRANSFORMATION STRESSES DURING QUENCH (NO BORE QUENCH)

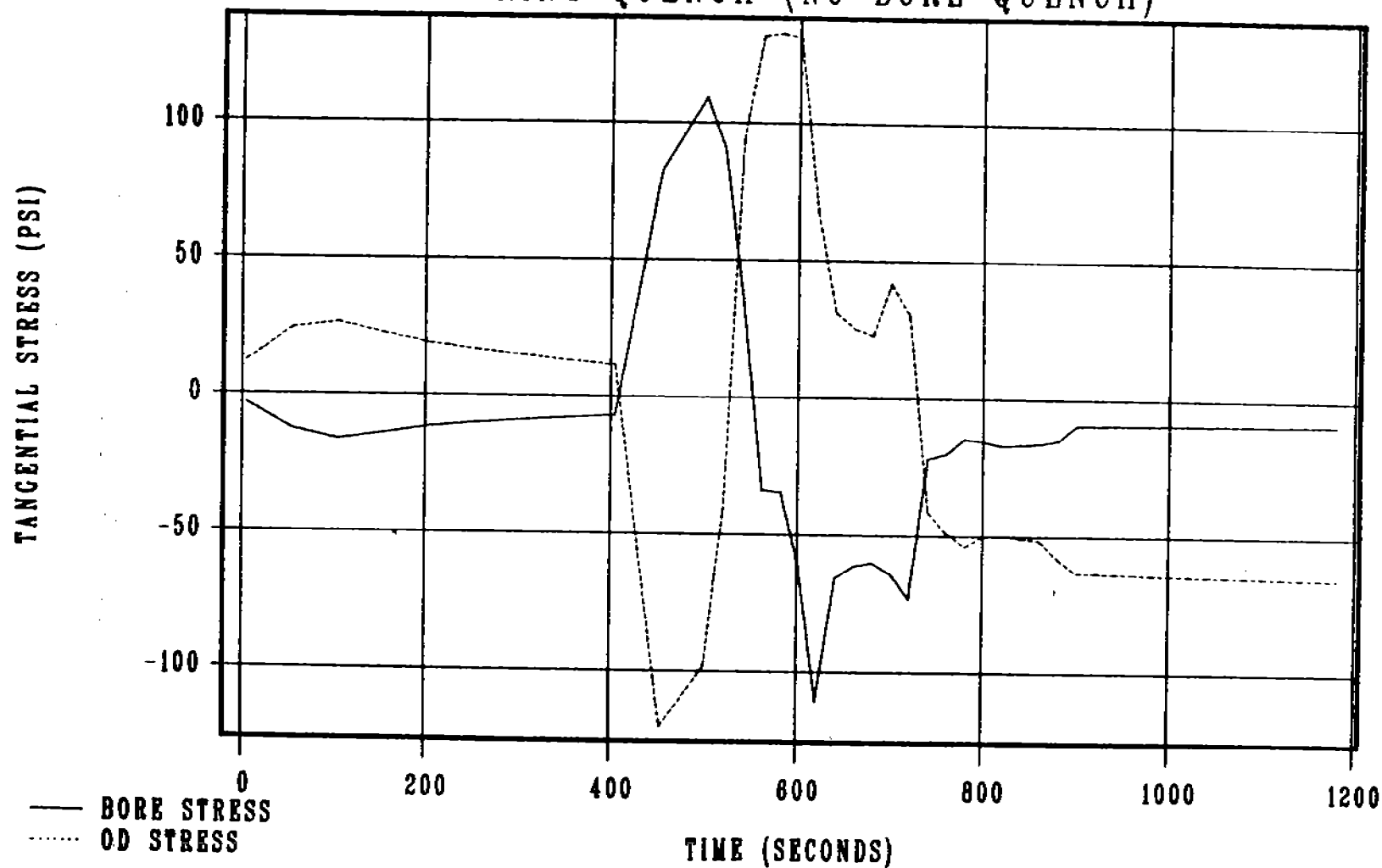


FIGURE 17.

BASED ON MODIFIED
TRANSFORMATION CURVE
(SLOW QUENCH) 12/31/84

FURTHER INVESTIGATION OF THE STABILITY OF DIFFUSION FLAMES NEAR EXTINCTION*

Y.S. Choi and G.S.S. Ludford
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca, NY 14853

ABSTRACT. For fixed L_F (the fuel Lewis number) equal to 1, strong dependence of the exchange of stability point on L_0 (the oxidant Lewis number) is found near extinction. On the other hand, with $L_0 = 1$ the effect of varying L_F on the location of the point is found to be minute.

1. INTRODUCTION. In a paper [1] at the last Army Conference, we described an investigation of both the near-extinction and near-ignition stability of diffusion flames. The emphasis was on Lewis-number effects and, for flames near extinction, the results were of a preliminary nature. The ones presented here, though not definitive, bring the story up to date. The investigation continues and we expect the remaining difficulties to be overcome soon.

2. GOVERNING EQUATIONS FOR NEAR-EXTINCTION ANALYSIS. The near-extinction steady states (corresponding to the bottom half of the S-response in figure 1) have been described in [2]. The perturbation temperature satisfies the differential equation

$$\frac{d^2 t_s}{d\xi^2} = -KL_0 L_F (k_1 \xi + k_2 - t_s)(k_3 \xi + k_4 - t_s) e^{t_s} \quad (1)$$

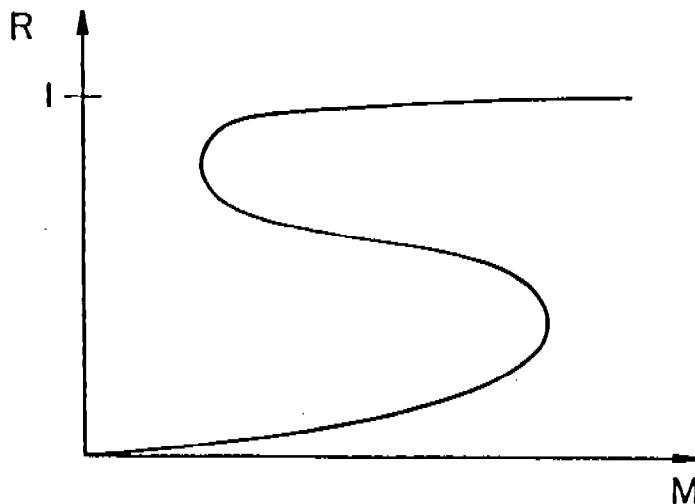


Figure 1. S-shaped response for chambered diffusion flame: R is the fraction of unburnt fuel and M is the injection rate.

*Supported by the U.S. Army Research Office.

and the boundary conditions

$$t_s = \begin{matrix} k_1 \xi - k_3 B + O(1) \\ k_3 \xi + k_1 C + O(1) \end{matrix} \quad \text{as } \xi \rightarrow \pm\infty. \quad (2)$$

Here k_1 is a known positive constant, $k_2 = PC$, k_3 is a known negative constant, $k_4 = QC$, where P, Q are known constants and B, C are unknown constants; the constant K is positive. The numerics determine exactly two solution for each K greater than a certain $K^*(L_0, L_F)$, exactly one for $K = K^*$ and none for $K < K^*$.

From these solutions the two responses

$$R = -\delta_a C,$$

where $\delta_a \ll 1$ is a (known) small positive constant, can be calculated for each $K > K^*$, thereby generating the middle and lower branches of the S-shaped response curve in figure 1. The bend is approached as $K \rightarrow K^*$ and remote parts of the two branches as $K \rightarrow \infty$.

The corresponding stability problem is

$$\begin{aligned} -\left[\frac{d^2 \phi_T}{d\xi^2} + \lambda \phi_T\right] &= K e^{t_s} (y_{0s} \phi_F + y_{Fs} \phi_0 + y_{Fs} y_{0s} \phi_T), \\ &= L_0^{-1} \frac{d^2 \phi_0}{d\xi^2} + \lambda \phi_0 = L_F^{-1} \frac{d^2 \phi_F}{d\xi^2} + \lambda \phi_F \end{aligned} \quad (3)$$

$$\phi_T(\pm\infty) = \phi_0(\pm\infty) = \phi_F(\pm\infty) = 0, \quad (4)$$

where

$$y_{0s} = L_0(k_1 \xi + k_2 - t_s), \quad y_{Fs} = L_F(k_3 \xi + k_4 - t_s)$$

are the mass fractions corresponding to the steady state temperature perturbation t_s . Here λ is the eigenvalue; if the spectrum has non-negative real part for a steady state t_s , that state is stable to the class of disturbance considered; otherwise it is unstable.

3. NUMERICAL RESULTS. In our previous paper [1], we reported preliminary results of an extinction analysis. Here, a more accurate calculation is being recorded.

For $L_0 = L_F = 1$, more careful computation shows agreement with Buckmaster. Nachman and Telifferro's result ([3]) i.e., change of stability occurs at the turning point of steady states.

With general L_0, L_F , we have to deal with 3 second-order simultaneous equations on a doubly infinite domain in order to extract the required eigenvalue. Computations show that although the steady-state response can generally be obtained to a higher degree of accuracy with a reasonable numerical infinity, the corresponding stability problem requires a much larger numerical infinity for convergence of eigenvalues. In some cases, we even have difficulty in getting the eigenvalue converged. (For small L_0 , for example 0.1, we even had difficulty in getting the steady-state response, because C is so small.)

From these numerical experiments, it is found that the real part of the smallest eigenvalue decreases as the numerical infinity is increased. Hence, once a negative eigenvalue is obtained on the lower branch, we may expect the location of the change of stability point to remain on the lower branch when more accurate computations are made.

With L_F fixed at 1, for very large L_0 the change of stability point is on the lower branch. As we gradually decrease L_0 to 1, the point moves to the turn. Further decrease of L_0 makes the change of stability point move back to the lower branch until a certain critical value is attained which we conjecture to be $(L_0)_{\text{crit}} = \log(1+Y_{0,1})/\log 2$ (cf. [2]). As we now decrease further still, the change of stability point moves towards the turn again. For L_0 about 0.13, it is very close to the turn. However, numerical difficulties prevent us from getting information for smaller values of L_0 .

On the other hand, with $L_0 = 1$ and L_F varied, the effect on the change of stability point is very minute, at least for the few cases that have been tested.

Re-examination of the numerical procedure is in progress. Some theoretical results and ideas have also been developed. We hope to give a better justification of the numerical results obtained in the near future.

REFERENCES.

- [1] Y.S. Choi, C. Laine-Schmidt and G.S.S. Ludford (1984). Numerical Investigation of the Stability of Diffusion Flames near Extinction and Ignition. Transactions of the Second Army Conference on Applied Mathematics and Computing (ARO Report 85-1), 225.

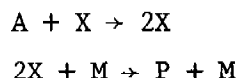
- [2] Y.S. Choi (1986) Chambered Diffusion Flames for Arbitrary Lewis Numbers.
Ph.D. Thesis, Cornell University.
- [3] J. Buckmaster, A. Nachman & S. Taliaferro (1983) The fast time instability of diffusion flames. Physica D #9, 408.

COMPLEX KINETICS IN FLAME THEORY*

G.S.S. Ludford and Richard Y. Tam
Department of Theoretical and Applied Mechanics
Cornell University
Ithaca, NY 14853

ABSTRACT. The Zeldovich-Liñán model with two-step kinetics is chosen to illustrate the complexities involved and the methods used for complex kinetics. A systematic procedure due to Fife and Nicolaenko characterizes each reaction by its power function, i.e. the part of the reaction source term that does not depend on the mass fractions. Of particular interest are the cross-over temperature T_c , at which the two power functions are equal, and its relation to the upper and lower bounds, T_u and T_ℓ , of the burnt temperature, which of course depend on the state of the fresh mixture. The possibilities $T_c < T_\ell < T_u$, $T_\ell < T_c < T_u$, $T_\ell < T_u < T_c$ lead to three different flame structures, in which the flame temperature is T_ℓ , T_c , T_u respectively. Once the flame temperature has been determined the analysis follows that in the familiar asymptotic treatment of one-step kinetics.

INTRODUCTION. The one-step kinetic model has served combustion modelling well. However, some important phenomena, e.g. flame quenching, are not reproducible by the one-step model. The natural extension is therefore to multi-step kinetics which involve intermediate products called radicals. It has long been known that free radicals can exert considerable influence on combustion processes. A simple two-step model that involves one intermediate species is the Zeldovich-Linan model:



consisting of a chain-branching (production) step and a chain-breaking (recombination) step. Here X is the radical, A the reactant and M a third body. The activation energy of the production step is very large, while that of the recombination step is small and thus taken to be zero. We shall use this model to illustrate a systematic method of evaluating complex kinetic schemes that is due to Fife and Nicolaenko (see Nicolaenko, 1985).

$$\begin{aligned} (1) \quad \frac{dY}{dx} - L^{-1} \frac{d^2 Y}{dx^2} &= - \nu_1 X Y e^{-\theta/T}, \\ \frac{dX}{dx} - K^{-1} \frac{d^2 X}{dx^2} &= - \nu_1 X Y e^{-\theta/T} - \nu_2 X^2, \\ \frac{dT}{dx} - \frac{d^2 T}{dx^2} &= q_1 \nu_1 X Y e^{-\theta/T} + q_2 \nu_2 X^2. \end{aligned}$$

*Supported by the U.S. Army Research Office.

Here T is the temperature, while X and Y are the mass fractions of radical and reactant (respectively); K and L are the Lewis numbers of the radical and reactant;

$$D_1 = D_1/M^2, \quad D_2 = D_2/M^2,$$

where D_1 and D_2 are the rate constants of the reaction steps; q_1 and q_2 are the proportions of the total heat released in the first and second steps of the reaction, so that

$$q_1 + q_2 = 1;$$

θ is the activation energy of the first step, that of the second step being taken zero. Given the parameters L , K , D_1 and D_2 , q_1/q_2 and θ , the problem is to determine the burning rate M for which these differential equations have a solution satisfying the boundary conditions

$$X, Y, T \rightarrow 0, \quad Y_f, T_f \quad \text{as } x \rightarrow -\infty,$$

$$X, Y, dT/dx \rightarrow 0 \quad \text{as } x \rightarrow +\infty.$$

The solution is sought in the limit $\theta \rightarrow \infty$.

POWER FUNCTIONS. Integrating the equations with respect to x from $-\infty$ to $+\infty$ and noting that the derivatives at both ends tend to zero, we get

$$Y_f = \alpha_1, \quad X_b = \alpha_1 - \alpha_2, \quad T_b - T_u = q_1 \alpha_1 + q_2 \alpha_2,$$

where

$$\alpha_1 = \int_{-\infty}^{\infty} D_1 X Y e^{-\theta/T} dx, \quad \alpha_2 = \int_{-\infty}^{\infty} D_2 X^2 dx.$$

Note that $0 \leq X_b \leq Y_f$ which at the two ends of the inequality yields

$$T_\ell = T_f + q_1 Y_f, \quad T_u = T_f + Y_f;$$

these two values of T_b depend on the fresh state. Of course, we have $T_\ell < T_* < T_u$, where T_* is the flame temperature.

The part of the reaction source term that does not depend on the mass fractions is called a power function, which we denote by

$$H_1 = D_1 e^{-\theta/T}, \quad H_2 = D_2.$$

The cross-over temperature is that temperature at which the power functions are equal, i.e.

$$D_1 e^{-\theta/T_c} = D_2 \quad \text{or} \quad T_c = \theta / (\ln D_1 - \ln D_2).$$

Thus, in contrast to the two values of the burnt temperature found above, the cross-over temperature depends on details of the reaction such as activation energy, rate constants, etc.

Consider

$$\frac{\alpha_2}{\alpha_1} = \frac{\int_{-\infty}^{\infty} D_2 X^2 dx}{\int_{-\infty}^{\infty} D_1 X Y e^{-\theta/T} dx} = \frac{\int_{-\infty}^{\infty} D_2 X^2 dx}{\int_{-\infty}^{\infty} D_1 X Y e^{-\theta/T} dx} = \frac{\int_{-\infty}^{\infty} H_2 X^2 dx}{\int_{-\infty}^{\infty} H_1 X Y dx}$$

and note that α_1 must be no smaller than α_2 (since X_b must be nonnegative). Suppose $D_1 > D_2$, so the curves of the power functions H_1 and H_2 intersect to define the cross-over temperature T_c as shown in fig. 1. In the limit $\theta \rightarrow \infty$, for which all the first reaction is concentrated at a flame sheet, there are three possible relations between the flame temperature T_* and the cross-over temperature T_c , in the discussion of which we shall use, \ll to denote exponentially smaller than.

- (i) $T_* < T_c$: Then $H_1 \ll H_2$ and the only way to satisfy the inequality $\alpha_2/\alpha_1 \leq 1$ is to take $X \ll 1$ everywhere. This implies $X_b = 0$ and $\alpha_2 = Y_f$, so that $T_* = T_b = T_u$. Recombination is completed within the flame sheet.
- (ii) $T_* > T_c$: Then $H_2 \ll H_1$ and $\alpha_2 \ll \alpha_1$; no restriction is imposed on the order of magnitude of X . Consequently, we have $\alpha_2 = 0$, from which follows $X_b = Y_f$ and $T_* = T_b = T_\ell$. Radicals remain at $x = +\infty$, and recombination takes place over exponentially long distances, i.e. downstream of what may be called the flame proper.
- (iii) $T_* = T_c$: H_2 being comparable (exponentially) to H_1 , the recombination reaction must go to completion on the x -scale, and consequently, $X_b = 0$, $\alpha_2 = Y_f$ and $T_* < T_b = T_u$.

THE FIVE POSSIBLE FLAME STRUCTURES. Now, there are three possible orderings of the cross-over temperature T_c and the burnt-temperature bounds T_ℓ and T_u (two additional limiting cases c can also be identified). The three possibilities may be characterized by the terms (a) fast, (b) slow, and (c) intermediate recombination. They occur in the following circumstances (see Figure 2).

- (a) $T_\ell < T_u < T_c$. Case (i) applies and the flame temperature has to be at its upper bound;

- (b) $T_c < T_\ell < T_u$. Case (ii) applies and the flame temperature has to be at its lower bound;
- (c) $T_\ell < T_c < T_u$. Assuming either case (i) or (ii) leads to a contradiction, so case (iii) applies and the flame has to be at the cross-over temperature.

These results are summarized by saying that T_* is driven as close as it can be to T_c . Determination of the flame temperature is a crucial step in solving the problem. The asymptotic analysis is now similar to that for the one-step reaction, and leads to T , X , Y profiles sketched in figure 2, as well as formulae for the burning rate. It is found that

$$M \propto D_1^{1/2} \theta^{-3/2} e^{-\theta/T_u} \text{ for fast recombination,}$$

$$M \propto D_1^{1/2} \theta^{-1} e^{-\theta/2T_\ell} \text{ for slow recombination,}$$

$$M \propto D_1^{1/2} \theta^{-2} e^{-\theta/2T_c} \text{ for intermediate recombination.}$$

The two limiting cases are:

- (d) $T_\ell < T_u = T_c$. For this fast/intermediate recombination, the profiles are similar to those for the fast recombination case except that the radical concentration is much larger (but still confined to the flame sheet).
- (e) $T_c = T_\ell < T_u$. For this intermediate/slow recombination, the profiles are similar to those for the slow recombination case except that the recombination is completed by $x = +\infty$.

CONCLUDING REMARKS. It is now clear that for $D_1 < D_2$, only (a) can occur. But, for $D_1 > D_2$ (as we have supposed above) there are five possibilities depending on the state of the fresh mixture.

Fife and Nicolaenko have considered a whole range of multistep kinetic schemes. Their method is currently applied to the 4-step model of Peters & Smooke (1985), which has considerably more complexity. Now, apart from the 24 different ways of ordering the four reaction rate constants D_1, D_2, D_3, D_4 , there will be many cross-over temperatures. Which of these the flame temperature is driven close to is not as clear as for the two-step case. But the problem may not be as complex as it first appears.

REFERENCES.

- B. Nicolaenko (1985). The mathematics of flames, to appear in *Physica D*.
- N. Peters & M.D. Smooke. (1985). Fluidynamic-chemical interactions at the lean flammability limit, *Combust. Flame* 60, 171-182.

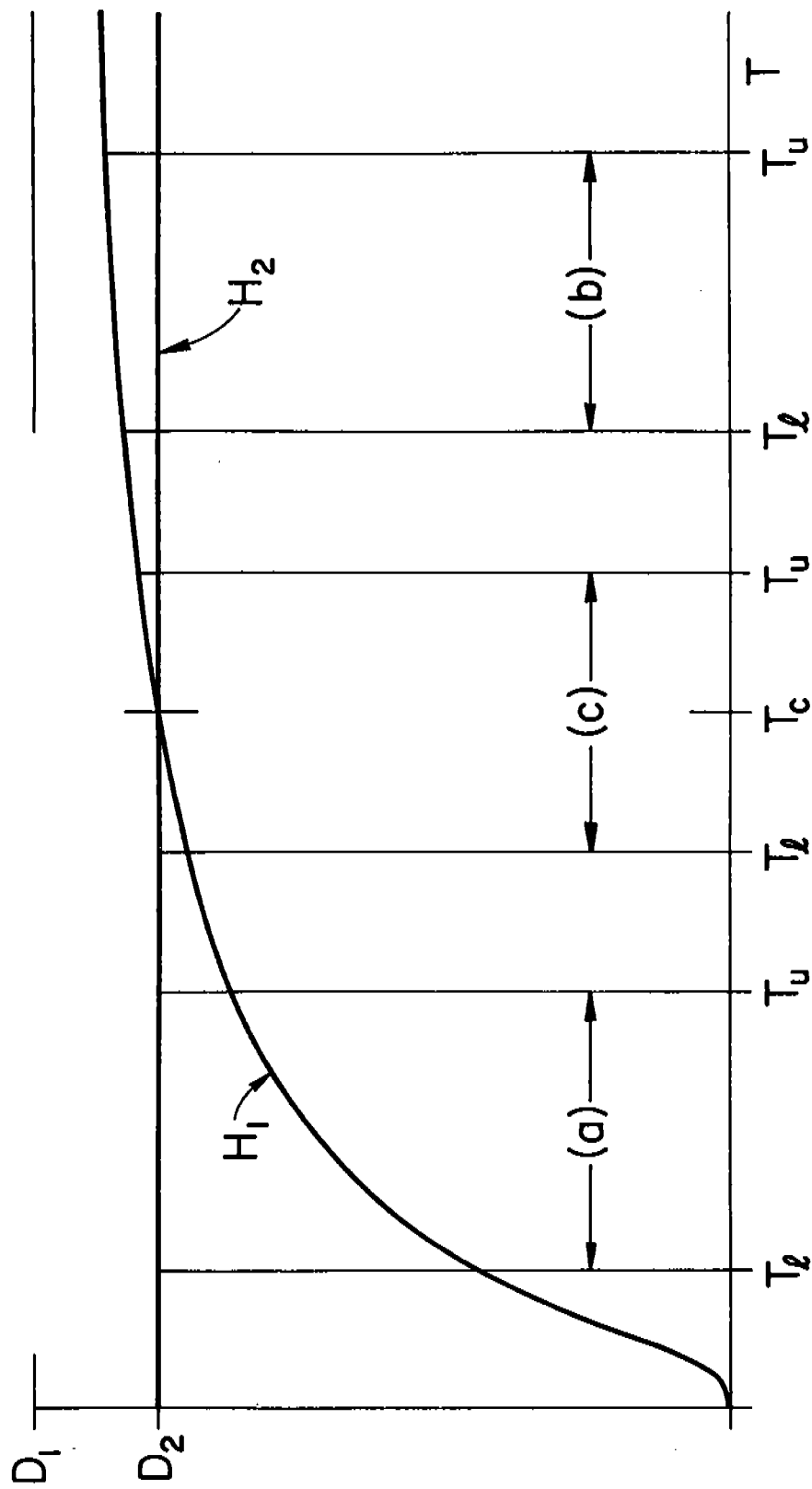


Figure 1

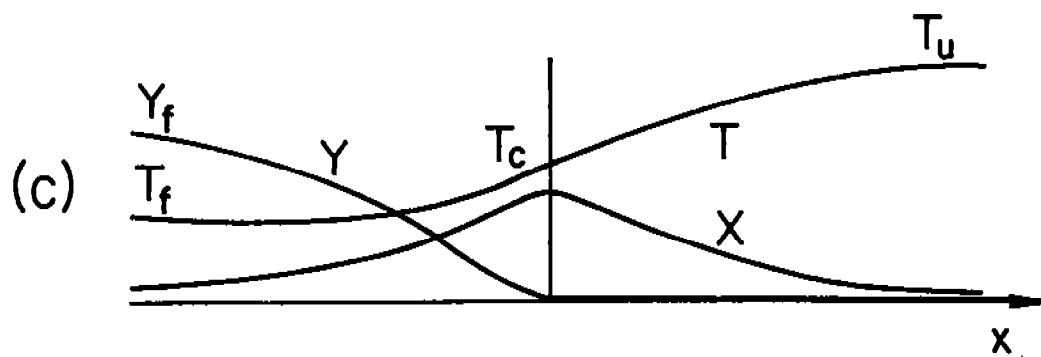
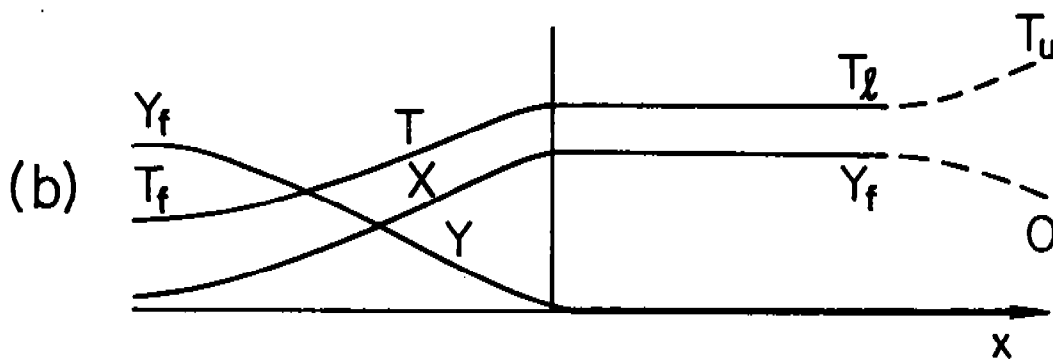
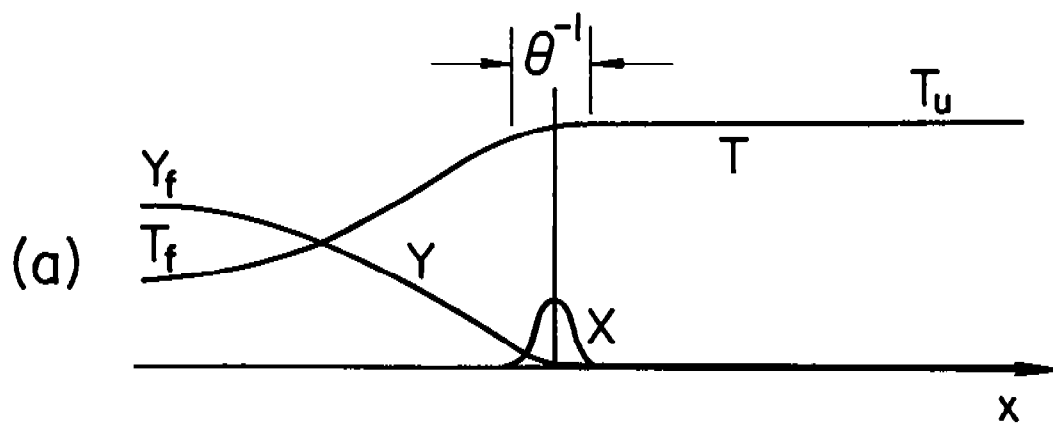


Figure 2

Applications of Front Tracking to Combustion, Surface Instabilities and Two Dimensional Riemann Problems: A Conference Report

Bruce Bukiet^{5,6}
*Carl L. Gardner*²
James Glimm^{1,2,3}
*John Grove*³
James Jones^{5,6}
Oliver McBryan^{1,2,6,7}

Courant Institute of Mathematics
New York University
New York, New York 10012

*Ralph Menikoff*⁴
*David H. Sharp*⁴

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

ABSTRACT

The method of front tracking is applied to problems involving curved detonation fronts, surface instabilities and two-dimensional Riemann problems. The detonation problems include detonation fronts with and without cylindrical symmetry; comparisons with one-dimensional models are made. The analysis of interface instabilities focuses on the compressible Rayleigh-Taylor instability of a supersonic accelerated contact discontinuity between two gases and the propagation of a supersonic slab jet. Theoretical notions for an S matrix theory for general multi-dimensional hyperbolic conservation laws and the numerical implementation of computer programs which solve certain two-dimensional Riemann problems are also discussed.

1. Introduction

Systems of non-linear conservation laws in n space dimensions

$$\mathbf{u}_t + \nabla \cdot \mathbf{F}(\mathbf{x}, \mathbf{u}) = 0 \quad (1.1)$$

1. Supported in part by the National Science Foundation, grant DMS-83-1229.
2. Supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U. S. Department of Energy, under contract DE-AC02-76ER03077.
3. Supported in part by the Army Research Office, grant DAAG29-83-K-0007.
4. Work supported by the U. S. Department of Energy.
5. Supported in part by the National Science Foundation, grant nos. MCS-82-07965 and MCS-83-01662.
6. Supported in part by the Army Research Office, grant DAAG-29-84-K0130.
7. Alfred P. Sloan Foundation Fellow.

where $u = u(x, t)$ and \bar{F} is a smooth function of the state $u \in R^r$ and the position $x \in R^n$ into R^r , are often used as first order approximations for many natural phenomena. Equations of this type occur in models in which external forces and higher order effects such as viscosity and heat conduction are neglected.

We are primarily concerned with the case where system (1.1) consists of the Euler equations for a compressible, inviscid, non-heat conducting gas. In this case:

$$u = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix} \text{ and } \bar{F}(u) = \begin{pmatrix} m \\ \frac{m \otimes m}{\rho} + p \\ \frac{m}{\rho}(E + p) \end{pmatrix}. \quad (1.2)$$

Here ρ is the density, m and E are the momentum and total energy per unit volume respectively and p is the thermodynamic pressure. These equations represent the laws of conservation of mass, momentum and energy respectively. The thermodynamic variables p , ρ and E are related by a caloric equation of state $p = p(\rho, E)$. For the case of a polytropic gas this relation is given by:

$$p = (\gamma - 1) \left[E - \frac{|m|^2}{2\rho} \right],$$

where γ is positive constant usually satisfying $1 < \gamma \leq \frac{5}{3}$.

Much progress has recently been made in adapting a front tracking method to the calculation of solutions of system (1.1) which contain discontinuities. In this method a one-dimensional grid is placed onto the discontinuity. Points on the tracked front are propagated by solving one-dimensional Riemann problems in the direction normal to the interface. This step provides the position of the tracked interface for the next time step. Tangential information is ignored during the normal propagation phase, so this step is followed by a update of the states on the new interface based on the component of (1.1) tangent to the interface. The positions of the new and old fronts together with their assigned states are used as boundary value data for the solution of the states off the front. A detailed description of this method can be found in [1].

The discontinuities supported by the Euler equations (1.2) are of two types, shocks and contact discontinuities. If combustion is considered, a third type of discontinuity, a combustion wave, may also occur.

In this paper we will report on recent progress which has been made by the authors and co-workers in modeling solutions in two space dimensions of the Euler equations for detonation waves, surface instabilities for non-combustion interactions and the numerical solution of certain two-dimensional Riemann problems. In addition recent theoretical work concerning a general theory of elementary waves and Riemann solutions for systems of hyperbolic conservation laws will also be discussed.

2. Multi-dimensional Riemann Problems and Elementary Waves

While the theory of hyperbolic conservation laws in one space dimension is highly developed, the corresponding theory for two or more space dimensions is not so well understood. Recent work has been devoted to the development of some basic notions which can be used in an S matrix theory for systems of hyperbolic conservation laws in more than one space dimension, see [2, 3, 4].

Let $\bar{A} = \frac{\partial \bar{F}}{\partial u}$ be the Jacobian matrix of \bar{F} . Equation (1.1) is said to be hyperbolic in a domain $D \subseteq R^{n+1}$ if the $s \times s$ matrix $\bar{A} \cdot \xi$ has real eigenvalues $\lambda_1, \dots, \lambda_s$ for all $(x, t) \in D$

and for all vectors ξ . If the eigenvalues λ_k are all distinct, the equation is said to be strictly hyperbolic. In the discussion which follows, hyperbolicity is assumed.

An S matrix theory is concerned with the large time asymptotic behavior of solutions to systems of equations for which system (1.1) is a first approximation. The leading order terms of these large time solutions are governed by the infinite scaling limit of the original system of equations. This scaling generally eliminates the higher order effects, yielding the system (1.1). It is assumed that any source terms in the original equation are of bounded extent. Under scaling these source terms will in general survive and go into a multiple of a delta function at the origin.

An S matrix is the product of two wave operators, the outgoing operator W^+ which gives the large positive time asymptotics and the incoming wave operator W^- which gives the large negative time asymptotics. Attention will be focused on the outgoing wave operator W^+ . The domain of W^+ usually taken to be the range of W^- . However, because of the occurrence of shocks in solutions of (1.1), this equation, when supplemented by the necessary entropy condition to separate physical from nonphysical waves, is not reversible. Thus W^- is not well defined. As a substitute the domain of W^+ will be restricted to scale invariant functions. Thus we consider solutions to the initial value problem for system (1.1) whose initial data is constant on rays through the origin. In some cases it will also be desirable to impose regularity conditions on the initial data as well. In one space dimension this is the well known Riemann problem and the problem of solving (1.1) with scale invariant data will be referred to as a multi-dimensional Riemann problem.

The notion of dimensionality in a Riemann problem is actually best described in terms of a co-dimension. A Riemann problem of co-dimension j is defined as the Cauchy problem for a system of conservation laws in d space dimensions in which the data is scale invariant in j dimensions and independent in the remaining $d - j$ dimensions.

The restriction to scale invariant data and the fact that equation (1.1) is itself scale invariant implies that a solution to a Riemann problem should be self-similar, that is, a function of $\frac{x}{t}$. This implies that a Riemann solution u of (1.1) will satisfy

$$-\frac{x}{t} \cdot \nabla u + \nabla \cdot \bar{F} = 0. \quad (2.1)$$

Such a solution u is completely determined by its values in the hyperplane $t = 1$ and by restricting our attention to this hyperplane time can be eliminated from the equation. Therefore a Riemann solution of (1.1) has in general one less degree of freedom than a general solution.

General solutions for Riemann problems are known in a few special cases. If the number of space dimensions is one, the system is strictly hyperbolic and each eigenvalue is either genuinely non-linear or linearly degenerate, then the classical paper of Lax [5] describes the solution of a Riemann problem with a small discontinuity as consisting of shocks, centered rarefactions and contact discontinuities. If the equation is scalar and in two space dimensions, solutions are known in the case where \bar{F} is of the form $f\bar{v}$, where f is a scalar valued function with at most one inflection point and \bar{v} is a constant vector [6, 7].

A central aspect of a scattering theory is that a source decomposes into some number of localized coherent waves, which then separate and propagate away from each other. These local disturbances are called d -dimensional elementary waves if the equation (1.1) is in d space dimensions. When d equals one these elementary waves include shocks, contact discontinuities and rarefaction waves. For $d > 1$ elementary waves are defined by the interaction of lower dimensional waves, for example when two shocks or a shock and a contact discontinuity collide. In many cases, such as the interaction of two shock waves, these waves will move with a definite velocity. Assuming that the original equation (1.1) is invariant under Galilean boosts one can then make a translation to a reference frame in which the wave is at rest, thus eliminating one more degree of freedom from the equation.

Understanding the structure of these elementary waves is crucial in describing the solution of a Riemann problem. In some special cases this structure is known. As mentioned above, in the case of one space dimension, hyperbolicity and a suitable type of convexity for the flux function, a theory of Lax describes these elementary wave as consisting of shocks, contact discontinuities and rarefaction waves which correspond to the eigenvalues of the differential of the flux function. If the equation is scalar, then a general theory of Oleinik [8] describes these waves in terms of the convex envelope of the associated flux function. Other special cases in which the structure of elementary waves is known include one space dimension polytropic gas dynamics [9], two space dimension polytropic gas dynamics [10], adsorption with a Langmuir isotherm [11], and water and polymer displacement of oil without adsorption [12, 13].

3. Two Dimensional Detonation Fronts

The method of front tracking has been applied to detonation waves in two-dimensional gas dynamics [14]. For problems which exhibit cylindrical symmetry comparisons can be made between the two-dimensional model and a one-dimensional model which exploits the symmetry and the agreement between these two methods is good. One finds that as the mesh of the computational grid is refined, the two-dimensional model converges linearly to the solution given by the model based on cylindrical symmetry.

Only a polytropic equation of state is considered, so the energy term E in (1.2) becomes:

$$E = \frac{p}{\gamma - 1} + \rho q + \frac{|m|^2}{2\rho}.$$

Here q is the the energy released by the chemical reaction that occurs across the detonation front. The Chapman-Jouguet model of detonation is used. In this model it is assumed that the reaction takes place instantaneously and that the reaction zone is infinitely thin.

If state 0 is the unburned gas and state 1 is the burned gas, the states on the two sides of the detonation are related by:

$$M^2 = \frac{p_1 - p_0}{\tau_0 - \tau_1}, \quad (3.1a)$$

where M is the mass flux across the front, and the Hugoniot relation:

$$\frac{\gamma_0 \tau_0 p_0}{\gamma_0 - 1} - \frac{\gamma_1 \tau_1 p_1}{\gamma_1 - 1} - (q_1 - q_0) = \frac{(p_0 - p_1)(\tau_0 - \tau_1)}{2}. \quad (3.1b)$$

In the case of a Chapman-Jouguet (CJ) detonation, the detonation wave moves at the local sound speed with respect to the gas behind it [15], and the behind state is completely determined by the ahead state and equations (3.1). When the combustion front is a strong detonation one additional parameter is necessary in order to specify the behind state from the ahead state.

A series of both strong and CJ detonation runs using grid sizes of 5 by 5, 10 by 10, 20 by 20, 40 by 40 and 80 by 80 have been made. The contact discontinuity behind the detonation front and the detonation front itself were tracked. Several of these runs were initially cylindrically symmetric and in these cases comparisons were made with a one-dimensional computation using the random choice method with 1500 points in the radial direction.

Figs. 3.1 - 3.5 present the results of a cylindrically symmetric computation in which the initial pressure ratio across the front is 100. The initial density is uniform and the gas releases 92.65% of its internal energy upon combustion. Figs. 3.1 and 3.2 show the positions

of the contact (inner quarter circle) and the detonation (outer quarter circle) at the beginning and end of the run respectively. Other figures include comparisons of pressure profiles and detonation wave speed (see Figs 3.3 and 3.4). The detonation wave speed error when calculated with respect to the one-dimensional code is less than 0.5%. Fig. 3.5 shows the convergence of the front tracking code to the one-dimensional code under mesh refinement.

In addition to cylindrically symmetric runs the front tracking code has also been applied to problems in which the initial interface is elliptical. If the initial states are the same as the ones described above, hot spots are produced behind the front in regions of small curvature and cold spots in corresponding regions of large curvature. The initial lengths of the major and minor axes are .3 and .15 for the detonation wave and .29 and .145 for the contact. Fig. 3.6 shows pressure contours and the waves just before the detonation wave breaks through the boundary on a 30 by 30 grid. The pressure is higher behind the flatter portion of the detonation wave than behind the rounder portion of the wave.

4. Supersonic Interface Instabilities

Interface fingering instabilities arise in a wide variety of physical contexts: inertial laser fusion, plasma fusion, instabilities of layers in stars, the instability of laser accelerated foils, and astrophysical jets. We have examined the compressible Rayleigh-Taylor instability of a supersonic accelerated contact discontinuity between two gases. The computed solutions exhibit a complicated set of nonlinear waves comprised of spike and bubble bow shocks, terminal shocks within the spike and bubble, Kelvin-Helmholtz roll-up of the spike tip, and contact surface waves. Detailed analysis is given in Ref. [16]. We have also studied the propagation of a supersonic slab jet in order to compare and contrast the jet wave structure with that of the supersonic accelerated surface.

A compressible gas interface which is accelerated by a shock (the Meshkov-Richtmyer instability [17,18]) is Rayleigh-Taylor unstable. If the interface is accelerated by a gravitational field, then the interface is unstable when the light fluid pushes the heavy. The important features of this instability can be modeled by imparting an initial kinetic energy to the contact discontinuity, which subsequently is allowed to advect freely. We assume that the problem is periodic in x with reflecting boundaries at the top and bottom of the computational region.

The problem can be parametrized in dimensionless units by the initial Mach number of the tip of the spike with respect to the heavy gas and by the initial density ratio ρ_b/ρ_a (b denotes the gas below the contact, a the gas above). The dimensional scales are set by the initial ambient pressure, perturbation wavelength, and initial ambient density of the heavy gas. The polytropic gas constant γ was set equal to 1.4.

An interesting set of wave structures emerges from this study. Figure 4.1 portrays the evolution of a Mach 2.8 density ratio 2 accelerated surface at $t = 0.4$. The flow is initially supersonic in both gases. The bow shocks in the lower gas have interacted to form a single shock, while the spike bow shock has interacted and joined with its periodic neighbors. The spike exhibits the characteristic Rayleigh-Taylor roll-up, and the contact shape indicates the presence of small-scale surface instabilities.

Just inside of the advancing spike a "terminal" shock wave is formed. The contact is advancing more slowly than the heavy gas inside of the spike. A shock wave, preceded by a rarefaction wave, is formed as the advancing heavy gas is slowed down to the contact velocity. A similar terminal wave is formed in the light gas inside of the advancing bubble.

This shock preceded by a rarefaction wave pattern can be clearly seen in the density cross section plots in both the supersonic accelerated surface run (Figure 4.1) and the supersonic jet run (Figure 4.2). Note that while the jet terminal shock propagates with the head of the jet beam, the accelerated surface terminal shocks are physical transients which decouple

from the late evolution of the contact instability.

The compressible Rayleigh-Taylor results differ from the incompressible case chiefly in the formation of the terminal compression waves and in the fact that the spike exhibits less roll-up. The accelerated surface problem differs from the gravitational instability in that the spike appears to attain a finite growth of aspect ratio approximately equal to 2 for our range of parameters.

A resurgence of interest in supersonic jets has been sparked by the observation of astrophysical jets emanating from the cores of active galaxies and by the subsequent success of theoretical [19] and computational analyses [20].

The evolution of a Mach 3, density ratio 10 slab jet at $t = 0.4$ is presented in Figure 4.2 [16]. The jet was initialized by injecting gas at a specified Mach number into an ambient gas at equal pressure. The boundary conditions are through-flow. The problem is parametrized by the Mach number of the jet with respect to the jet gas and the density ratio of jet to ambient gas. γ was set equal to 5/3. The results apply to jets from laboratory to astrophysical scales since the problem is independent of length scale.

The jet beam in our 80x120 grid computation is 5 grid blocks wide, while the beam is 20 grid blocks wide in the 160x300 grid computation of Norman, Smarr, and Winkler [20].

The density contour and cross-section plots in Fig. 4.2 indicate the presence of a bow wave (the flow is subsonic in the ambient gas) and of a terminal shock near the head of the jet beam, preceded by a rarefaction wave. This terminal shock system may explain the observed hot spots terminating astrophysical jets [20]. The contact shape displays the large scale Kelvin-Helmholtz roll-up of this jet, and the development of two-dimensional pinch waves.

The fact that we get reasonable results with a beam 5 grid blocks across illustrates one of the advantages of the front-tracking method. By placing additional degrees of freedom around the tracked contact, the method is able to resolve the solution globally with fewer degrees of freedom than required by conventional finite difference methods. The importance of this feature of the method will become apparent when the statistical regime of multiple fingers is considered.

5. Numerical Implementation of Elementary Waves

Further work on the development of computer code for modeling elementary waves is in progress. Previous papers [10,1] have reported the numerical implementation in the front tracking code of the elementary waves known as regular reflection and single Mach reflection, this section will discuss the case of shock and contact discontinuity interactions.

The simplest model of a shock and contact discontinuity interaction consists of an incident shock wave colliding with a contact discontinuity separating two different gases. The local result of this interaction is a configuration we call a diffraction node. A diffraction node consists of the incident shock wave, the contact discontinuity into which the incident shock collides, a reflected wave which is either a shock or a centered rarefaction wave and a deflected contact surface behind the incident shock. The model supposes that locally all of the shock or contact waves can be assumed to be straight, and that the solution in a neighborhood of the node is piecewise constant except for the possible reflected centered rarefaction wave. It is assumed that the point of intersection of these waves moves with a definite velocity and thus the interaction can be studied in a frame of reference in which the node is at rest. The description of a diffraction node then consists of the states in a neighborhood of the node together with the angles at which the waves at the node intersect.

In a dynamic model it is necessary to calculate the transformation to the steady frame of the node. This is equivalent to finding the velocity of the node in the given reference frame. This velocity can be approximated by propagating the incident shock and the contact

discontinuity into which it collides for one time step ignoring their interaction. The intersection of the two propagated curves is then used as the updated node position from which the node velocity can be calculated. Once the node velocity is known, the transformation to the steady frame of the node is performed. If one assumes that the data in front of the incident shock on both sides the contact discontinuity is known, and the strength of the incident shock is given, then the configuration around the diffraction node in the steady frame can be found by the intersection of shock polars in the pressure turning angle space, see Henderson [21]. Finally the configuration is translated back to the original frame of reference.

Figure 5.1 shows the result of the interaction of a planar shock wave colliding with a sinusoidally perturbed contact discontinuity. The incident shock is in air ($\gamma = 1.402$) and the contact surface separates air from sulphur hexafluoride ($\gamma = 1.092$). This interaction is known as a fast-slow interaction since the sound speed in air is greater than that in sulphur hexafluoride. The initial shock has a pressure behind to pressure ahead ratio of 100. It is interesting to note the extreme proximity of the transmitted shock and the deflected contact discontinuity near the node. We were quite pleased with the front tracking code's ability to resolve a configuration with such close curves. The rectangular mesh used for this run was 20×20 , and the separation between the transmitted shock and deflected contact is on the order of one tenth of a mesh block for a large portion of the computational region. Furthermore the transmitted wave lies on the sulphur hexafluoride side of the contact discontinuity and the value of gamma for this gas is so close to one as to make the resolution of waves on moderate sized grids difficult for most finite difference methods. We suspect that without front tracking one would need a rectangular mesh more than ten times as fine in each linear dimension in order to resolve both the deflected contact and the transmitted wave.

6. Bifurcations and Wave Interactions

One of the principal difficulties in any front tracking code is the handling of bifurcations and interactions in the tracked interface. Examples of these include the transition from regular to Mach reflection and the passing of waves through computational boundaries. Interactions of both of these types have been either fully or partially implemented in our front tracking code.

Figure 6.1 shows a planar shock wave incident upon a ramp. The problem is initialized with the shock normal to the wall. When the ramp is reached, a bifurcation occurs, in this case to a regular reflection. At a later time the point of regular reflection reaches the top of the ramp. At this point the regular reflection node lifts off the wall producing a Mach type reflection.

In figures 6.2 the front tracking code is used to follow the development of a compressible Kelvin-Helmholtz roll-up [1, 22]. Initially two gases of equal pressure and temperature but moving in opposite directions are separated by a slip discontinuity. This slip surface is given an initial perturbation which causes it to roll-up. The boundary conditions at the sides of the computational rectangle are periodic. As the surface rolls up portions of the surface cross the periodic boundaries. Any section of the surface which propagates past a periodic boundary is disconnected from the original curve and reinstalled periodically shifted to the opposite side of computational rectangle. A linking between the periodically connected curves is maintained so that periodic boundary conditions are enforced. The visual effect is that as one section of the slip surface propagates out of the computational window, we see the corresponding portion of the periodic neighbor moving into our picture.

Not only is the interaction of curves with computational boundaries of interest, but the interaction of nodes as well. Fig. 6.3 shows a diffraction node propagating past a computationally passive boundary. This problem is supersonic and the signals from the exterior of the right hand boundary are sufficiently weak that their influence on the solution can be

ignored. Thus the problem of node passing through such a boundary is simply a matter of identifying those curves at the exiting node which leave and those which remain. Exiting curves are deleted and the remaining curves are separately installed on the boundary. The main difficulty here is dealing with the numerical degeneracies which occur because very short curves are produced when the node first crosses the boundary.

References

1. I-L. Chern, J. Glimm, O. McBryan, B. Plohr, and S. Yaniv, "Front Tracking for Gas Dynamics," *J. Comp. Phys.*, vol. To appear.
2. James Glimm, "Elementary Waves and Riemann Solutions: Their Theory and Their Role in Science," *DOE Research and Development Report DOE/ER/03077-249*, 1985.
3. James Glimm and D. H. Sharp, "Elementary Waves for Hyperbolic Equations in Higher Space Dimensions: An Example from Petroleum Reservoir Modeling," *DOE Research and Development Report DOE/ER/03077-248*, 1985.
4. James Glimm and D. H. Sharp, "An S Matrix Theory for Classical Nonlinear Physics," *Foud. of Physics*, vol. To appear.
5. P. Lax, "Hyperbolic Systems of Conservation Laws II," *Comm. Pure and Appl. Math.*, vol. 10, pp. 537-566, 1957.
6. W. B. Lindquist, "Construction of Solutions for Two Dimensional Riemann Problems," *Adv. Hyp. PDE's.*, (To Appear).
7. W. B. Lindquist, "The Scalar Riemann Problem in Two Spatial Dimensions: Sufficiency Condition for Piecewise Smoothness of Solutions and its Breakdown," *J. Math. Anal.*, SIAM, (To Appear).
8. O. A. Oleinik, "Discontinuous Solutions of Non-Linear Differential Equations," *Uspehi Mat. Nauk*, vol. 12, pp. 3-73 (1957). English transl., Amer. Math. Soc. Transl. Ser. 2, vol. 26 (1963), 95-172
9. J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer Verlag, New York, 1983.
10. J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv, "Front Tracking and Two Dimensional Riemann Problems," *Adv. in Appl. Math.*, vol. To appear.
11. F. Helfferidge and G. Klein, *Multicomponent Chromatography: Theory of Interference*, Marcel Dekker, New York, 1970.
12. B. Keyfitz and H. Kranzer, *J. Diff. Eqs.*, vol. 27, p. 444, 1978.
13. E. Isaacson, *J. Comp. Phys.*, vol. To appear.
14. B. Bukiet, "Application of Front Tracking to Two Dimensional Curved Detonation Fronts," *To appear*.
15. R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, p. 212, Springer Verlag, New York, 1948.
16. C. L. Gardner, "Compressible Rayleigh-Taylor Instability of Supersonic Accelerated Surfaces," *To appear*.
17. R. D. Richtmyer, "Taylor Instability in Shock Acceleration of Compressible Fluids," *Comm. Pure and Appl. Math.*, vol. 13, p. 297, 1960.
18. E. E. Meshkov, *Izv. Akad. Nauk SSSR, Mekh. Zhidk. Gaz.*, vol. 5, p. 151, 1969.
19. R. D. Blanford and M. Rees, *Mon. Roy. Astr. Soc.*, vol. 169, p. 395, 1974.
20. L. L. Smarr, M. L. Norman, and K-H.A Winkler, *Physica*, vol. 12D, p. 83, 1984.

21. L. F. Henderson , "The Refraction of a Plane Shock Wave at a Gas Interface," *J. Fluid Mech.* , vol. 26, p. 607, 1966 .
22. B. Plohr, J. Glimm, and O. McBryan, "Applications of Front Tracking to Two-Dimensional Gas Dynamics Calculations," in *Lecture Notes in Engineering Vol. 3*, ed. J. Chandra and J. Flaherty, p. 180, Springer Verlag, New York, 1983.

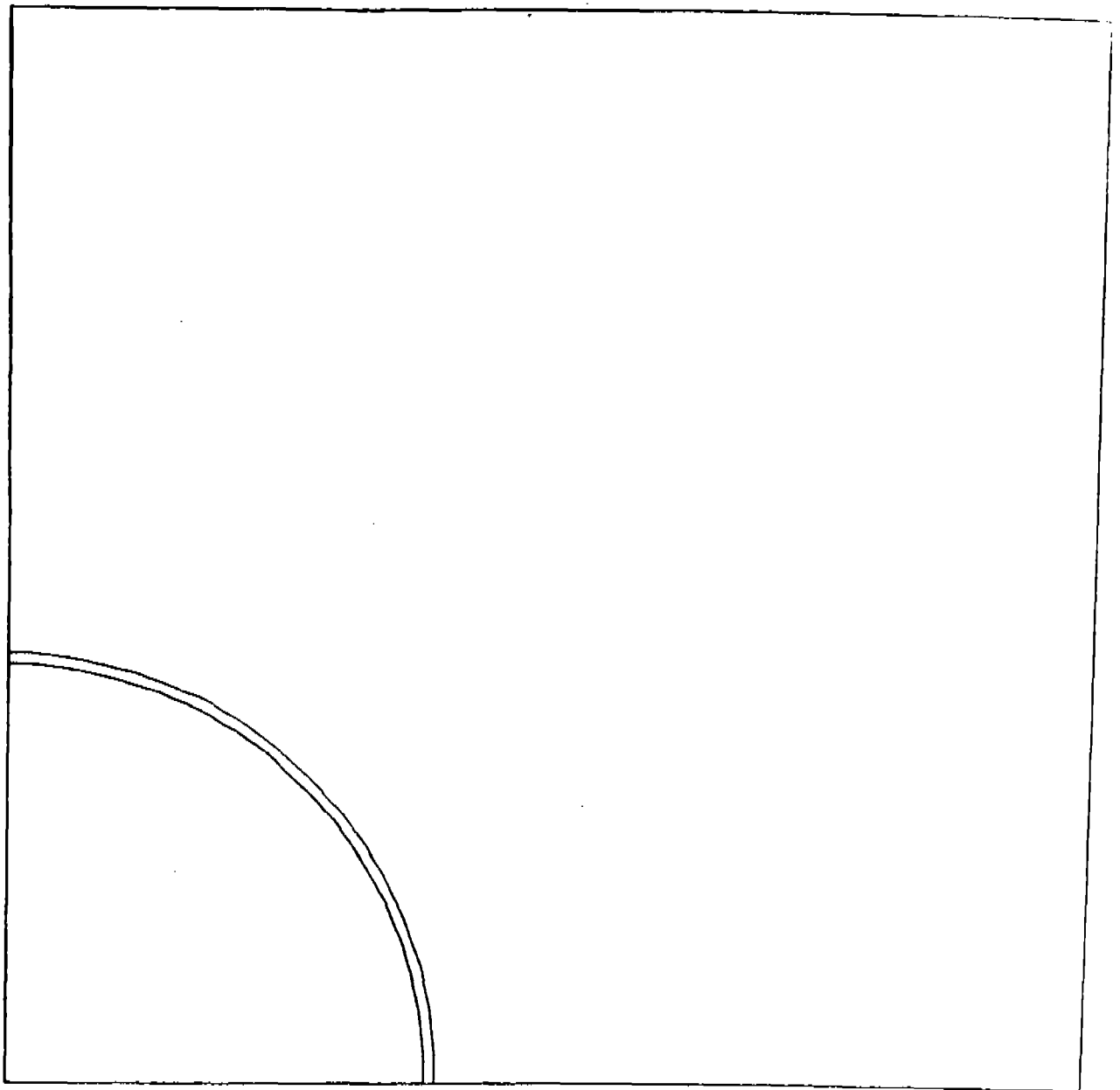


Fig. 3.1

The initial strong detonation wave (outer circle) and the contact discontinuity for a cylindrically symmetric computation. The initial conditions are uniform density, zero velocity, and a circular pressure discontinuity at radius .2, with ratio inside to outside of 100. The heat released upon combustion is 92.65% of the internal energy of the unburned gas. The initial position of the contact is radius .195. The initial state between the waves is that behind a planar detonation wave with the above initial data.

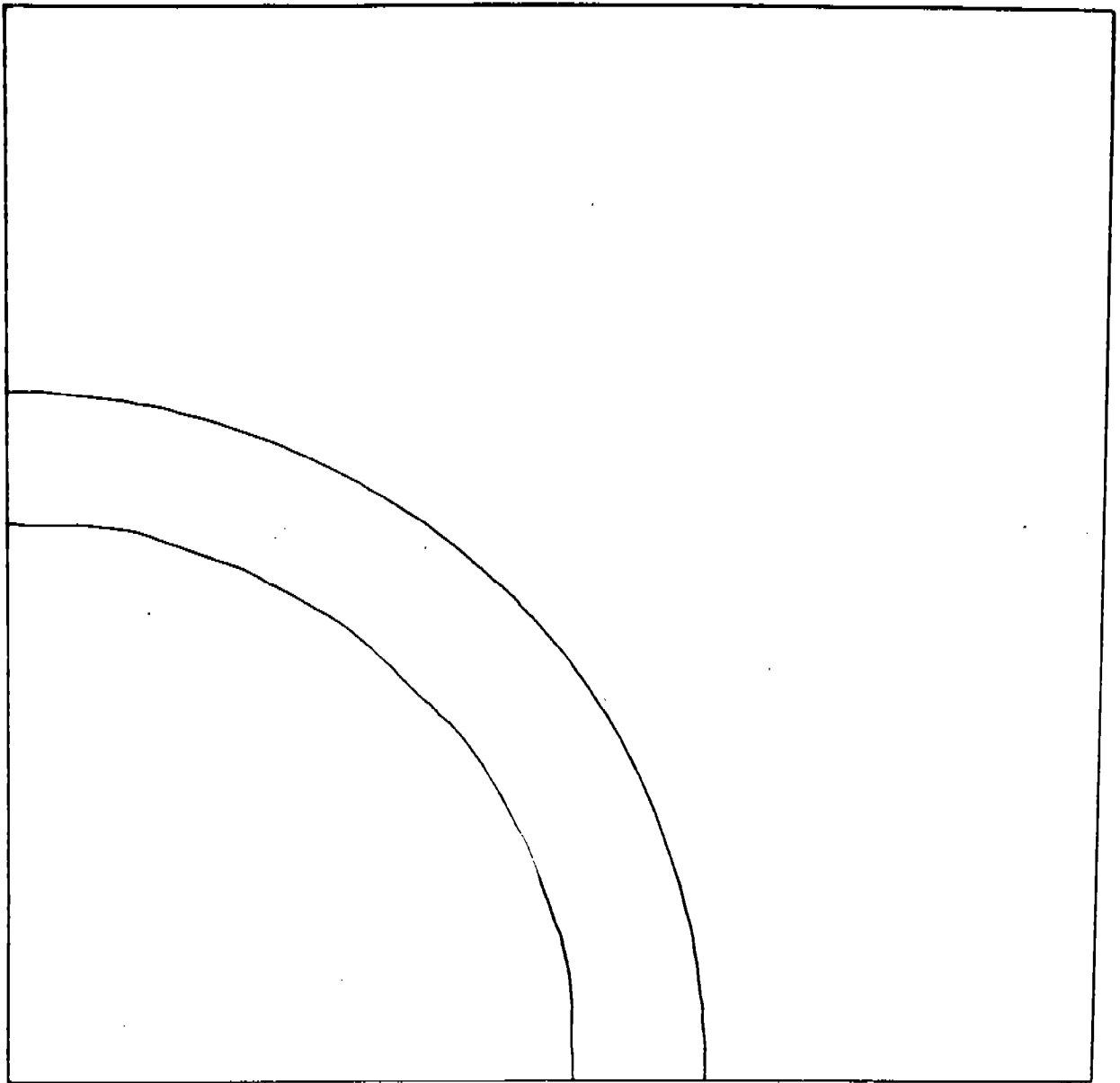


Fig. 3.2

The detonation wave and the contact at the time step analyzed in Figs. 3.3 and 3.4. The detonation wave now has radius approximately .36.

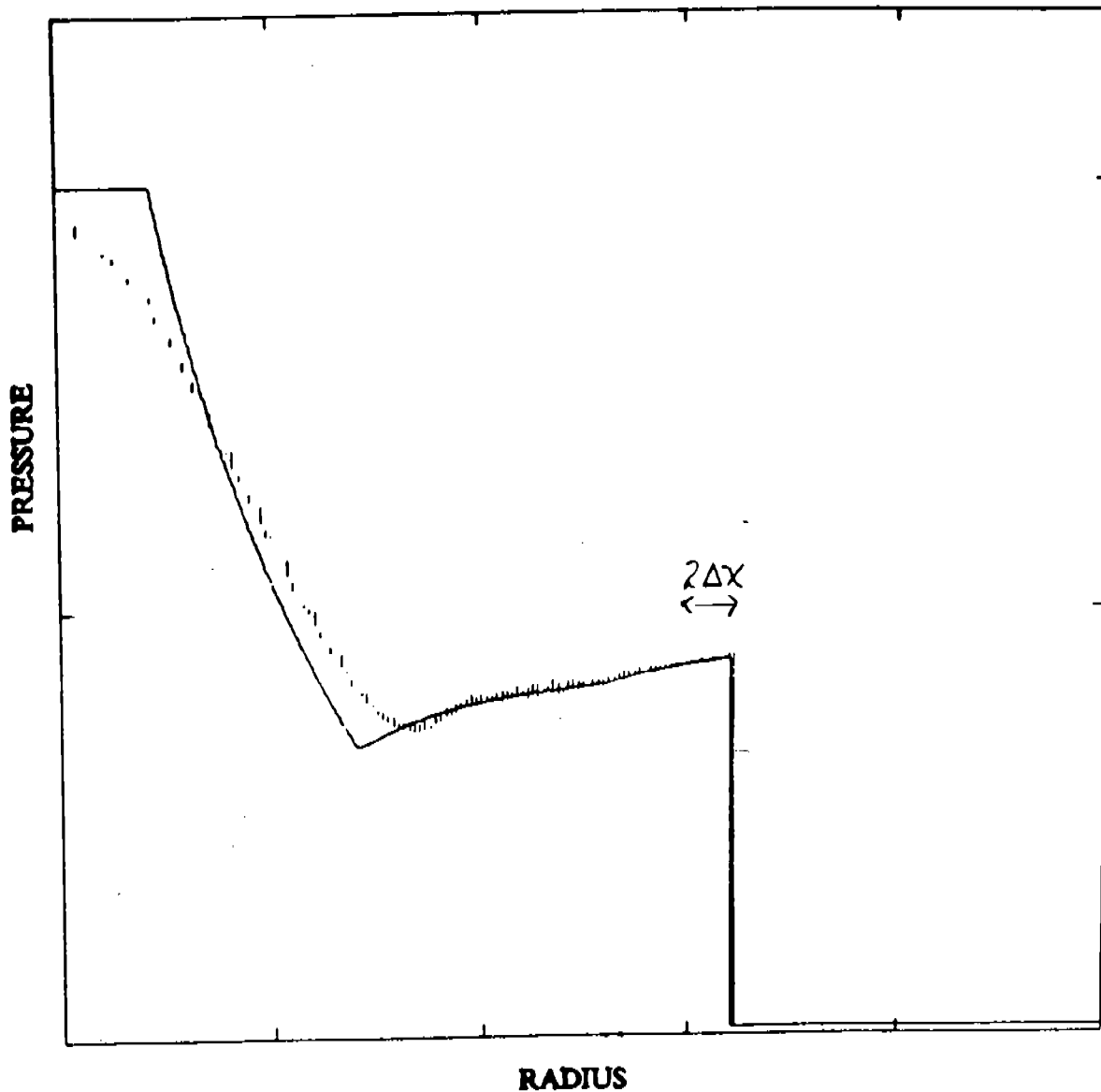


Fig. 3.3

A plot of pressure vs. radius corresponding to Fig. 3.2 is shown. The solid curve shows the results obtained by the one-dimensional random choice computation. The vertical lines represent the range of pressure values in the two-dimensional front tracking solution at a fixed radius as the angle varies on a 40 by 40 grid. Thus, the vertical lines show the angular dependence in the solution.

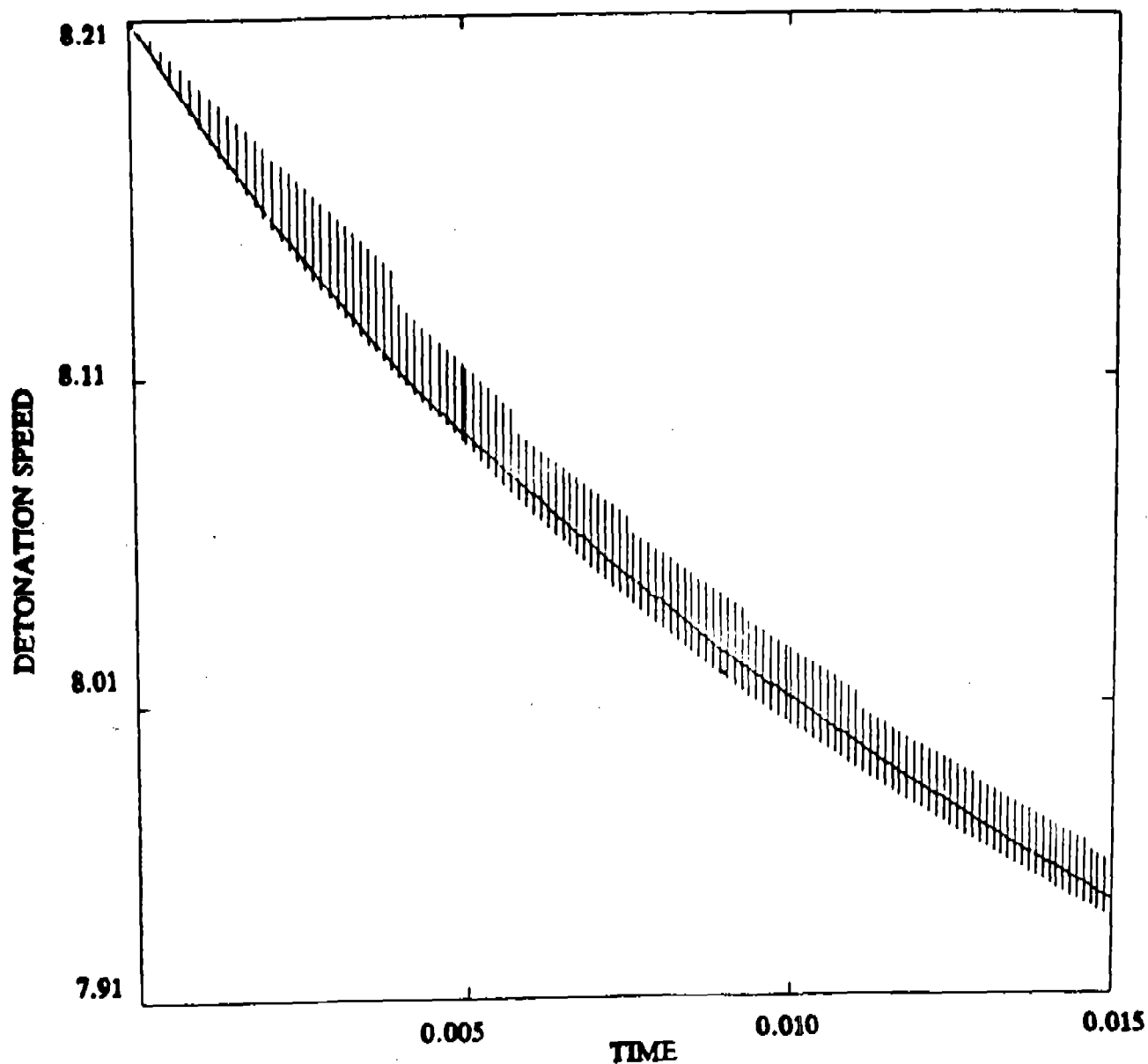


Fig. 3.4

A plot of detonation speed vs. time for the computation on a 40 by 40 grid. The solid curve shows the speed of the detonation wave in the one-dimensional calculation. The vertical lines represent the range of values of the speed of the detonation in the two-dimensional calculation. The maximum error ($\max \frac{|U_{2d} - U_{1d}|}{U_{1d}}$) is less than 0.5%.

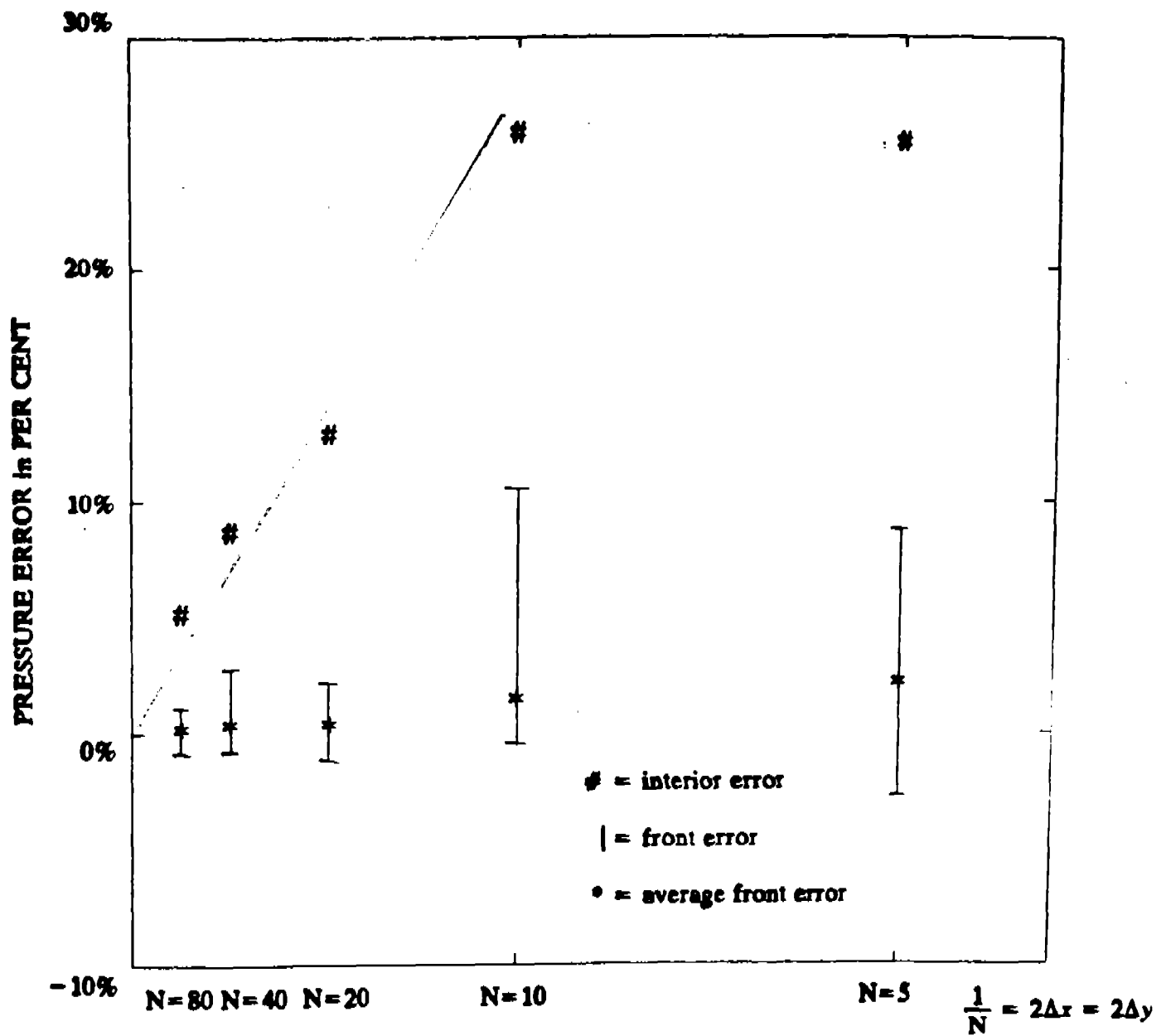


Fig. 3.5

Convergence of the front and interior schemes. The pressure errors in the interior and at the front are shown for $N \times N$ grids at the time indicated by Fig. 3.2. The # signs represent the interior error, where

$$\text{Interior Error} = 100\% \times \frac{\int_0^5 \int_0^5 |P_{2d} - P_{1d}| dx dy}{\int_0^5 \int_0^5 P_{\text{initial data}} dx dy}$$

The front error (error bars) gives the range of the errors at the front, defined as

$$\text{Front Error} = 100\% \times \frac{P_{2d} - P_{1d}}{[P]},$$

where $[P]$ is the pressure jump at the front in the one-dimensional computation at the same time. The asterisks represent the error of the average pressure behind the front, namely

$$\text{Front Error(average pressure)} = 100\% \times \frac{P_{2d \text{ average}} - P_{1d}}{[P]}.$$

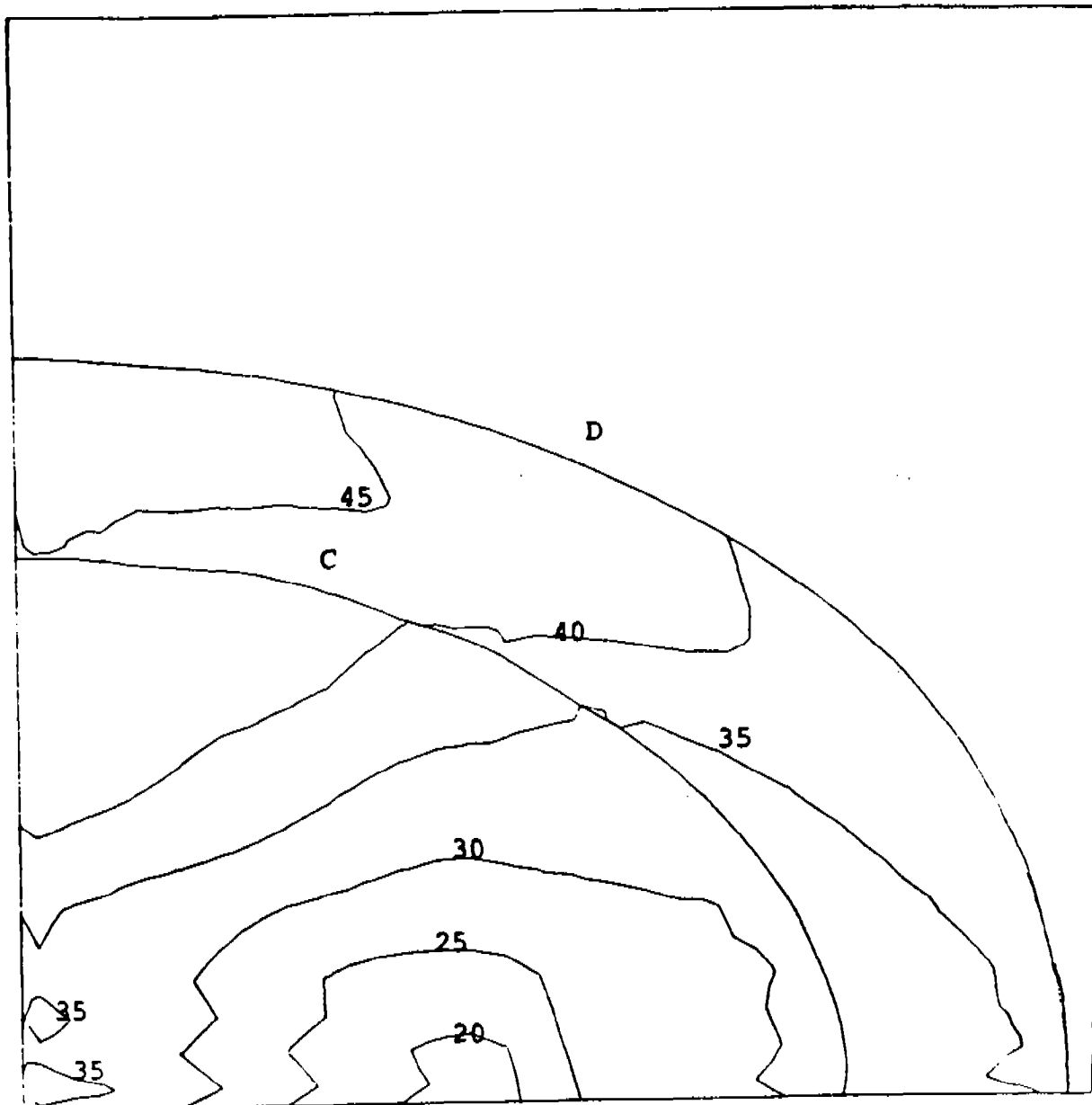


Fig. 3.6
Pressure contours are shown for a computation of an elliptical expanding detonation on a 30 by 30 grid. Also shown are the detonation wave (D) and the contact (C).

center-line density

edge density

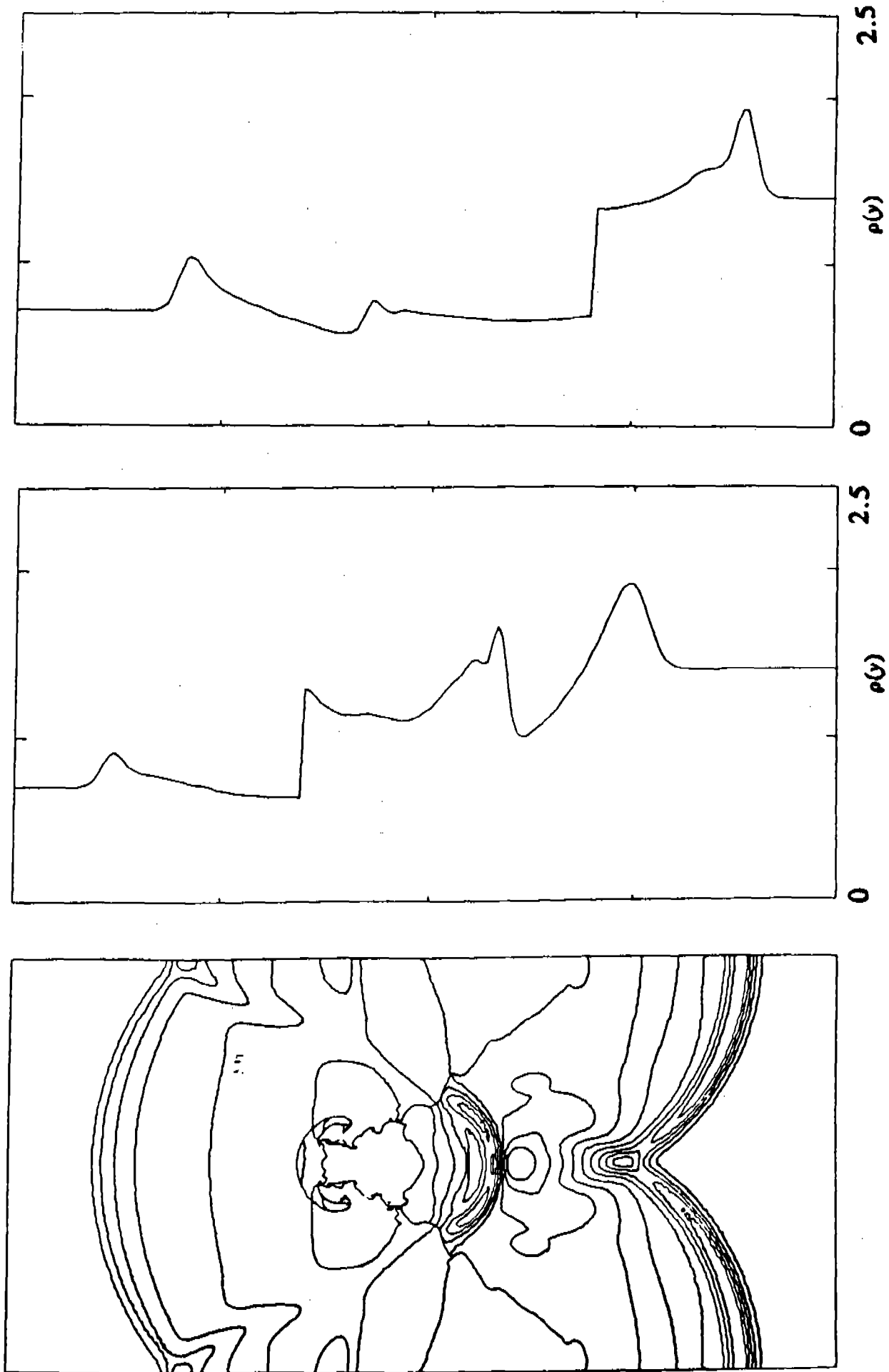


Fig. 4.1
 $M = 2.8$, $D = 2$ density contours and density cross sections at $t = 0.4$. 80×160 grid.

center-line density

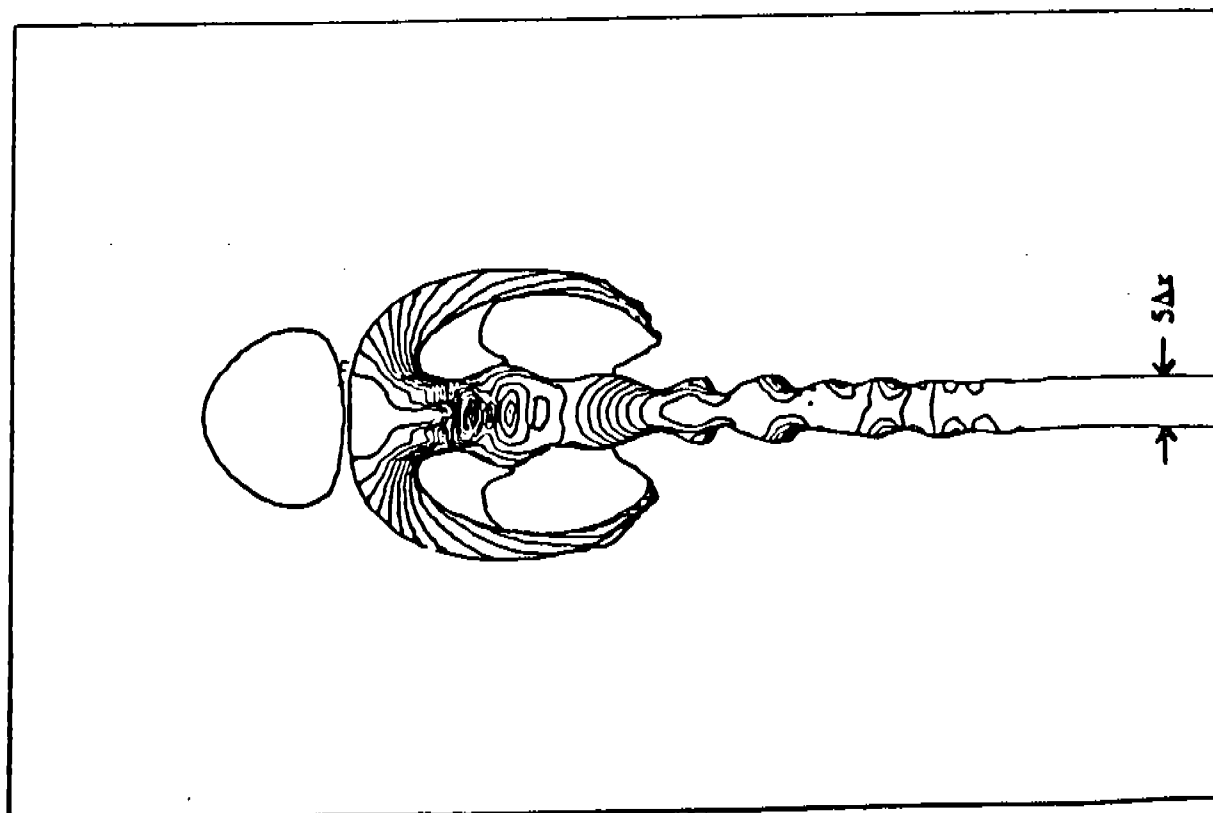
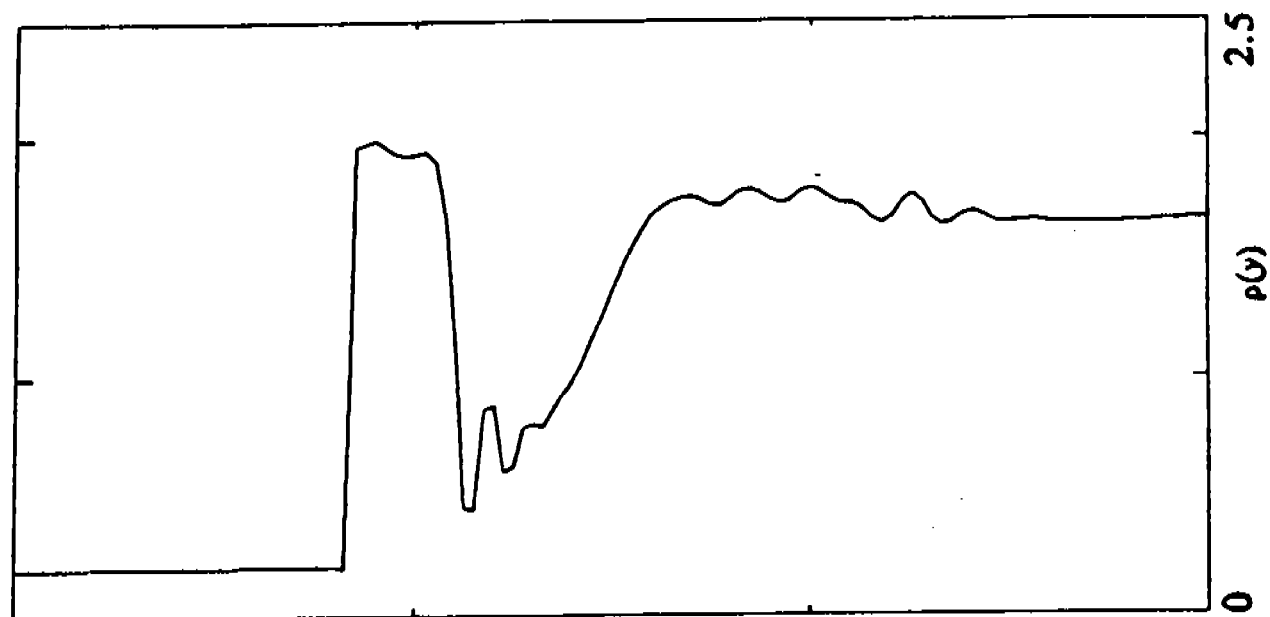


Fig. 4.2
Density contours and density cross section for $M = 3$, $D = 10$ jet at $t = 0.4$. 80x120

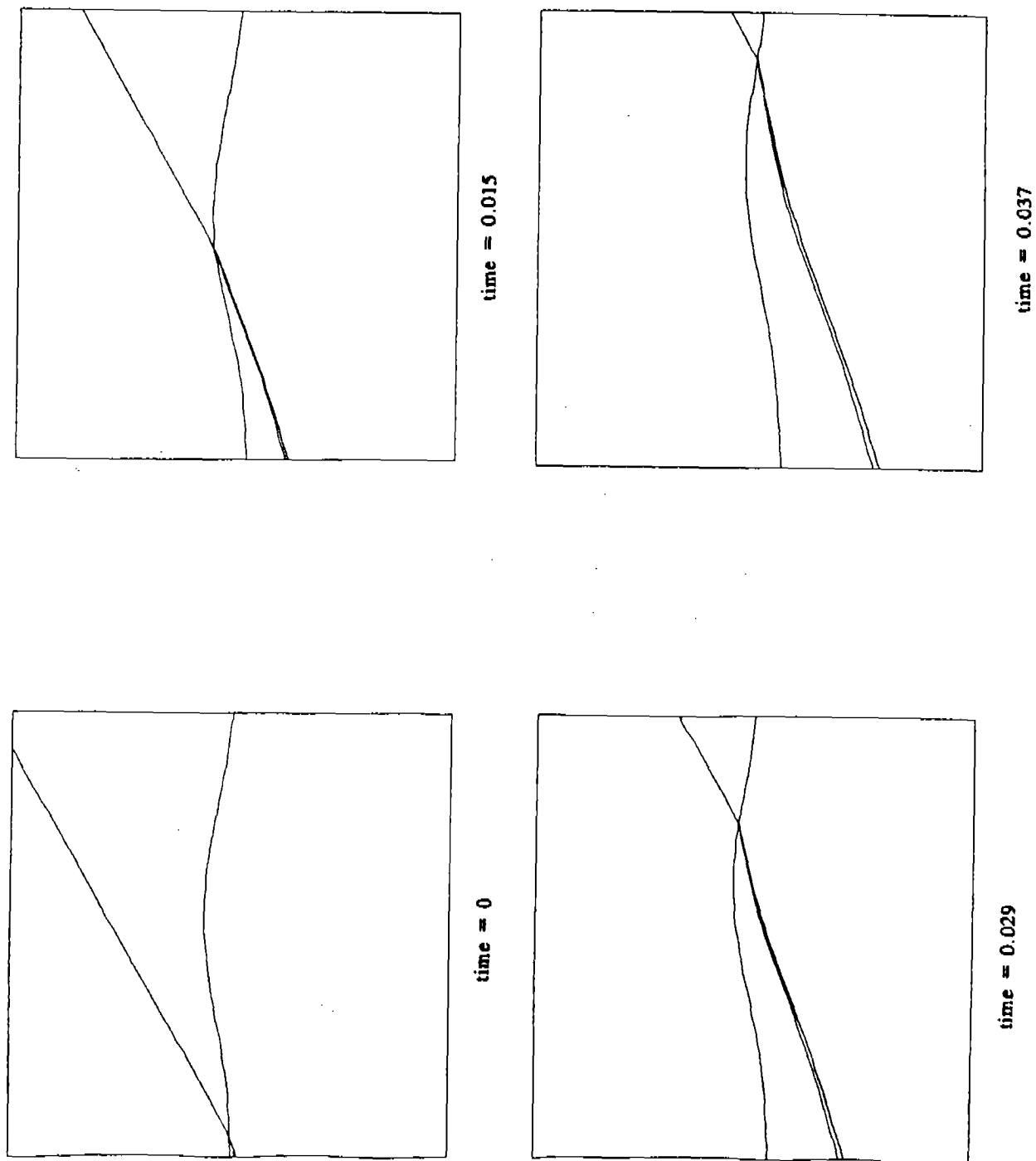


Fig. 5.1
The interaction of a shock wave with a contact discontinuity producing reflected and transmitted shock on a 20 by 20 rectangular grid.

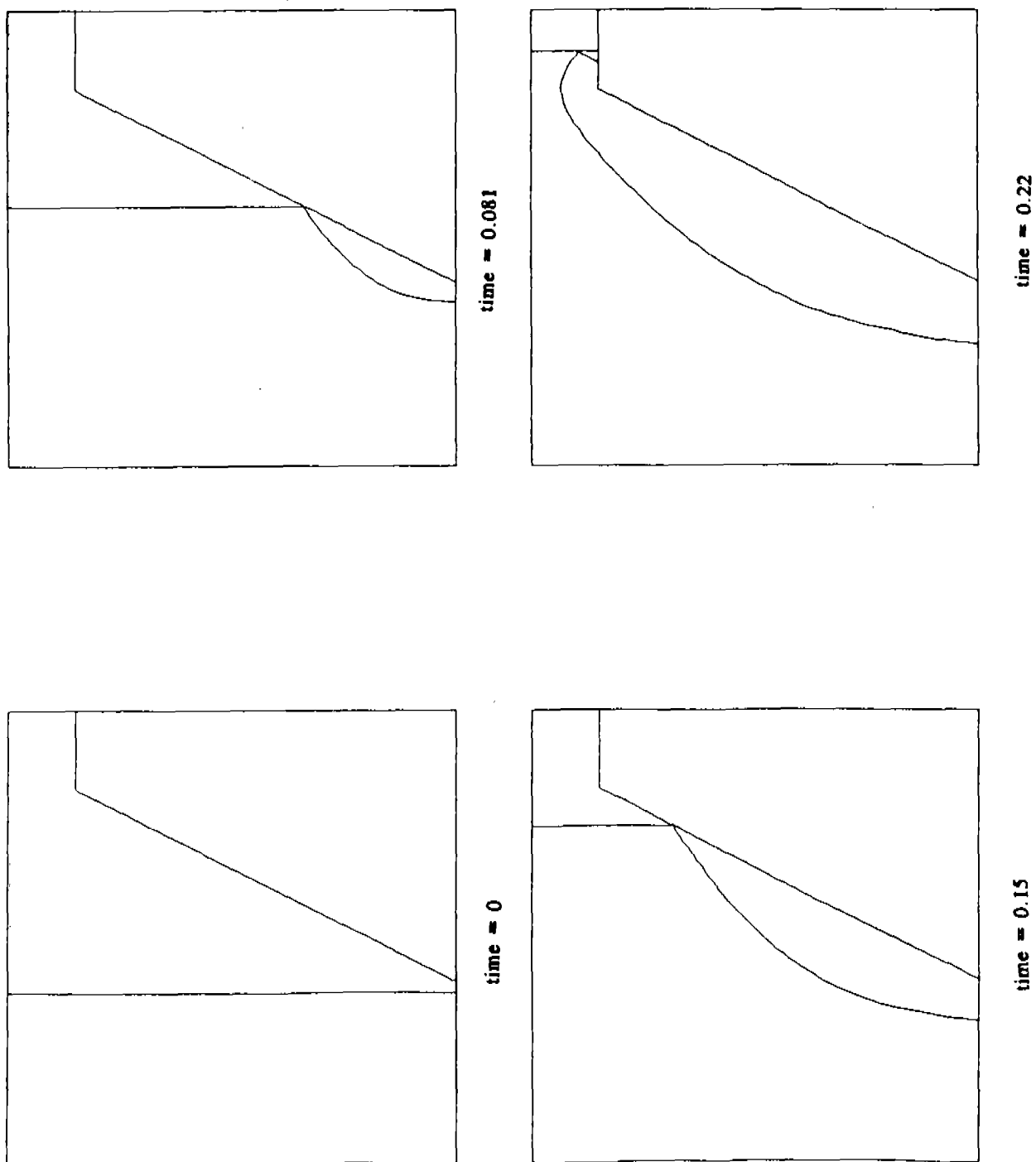


Fig. 6.1

A shock incident upon a ramp. Bifurcation to a regular reflection occurs when the shock reaches the ramp. When the regular reflection node reaches the top of the ramp a bifurcation to a Mach type reflection occurs. The grid here is 30 by 30.

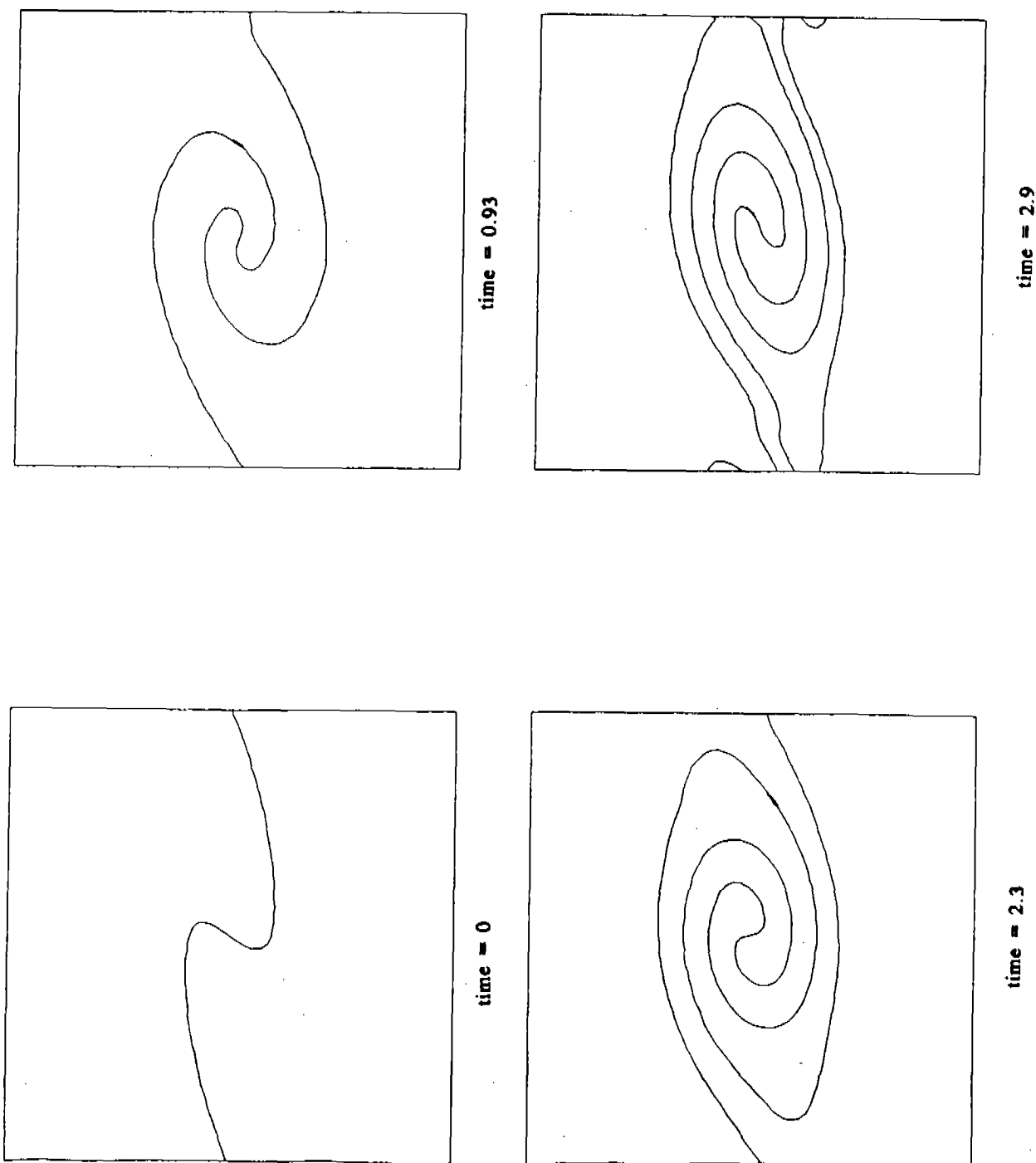


Fig. 6.2

Compressible Kelvin-Helmholtz roll-up. The computation is on a 40 by 40 grid. When interior curves cross the periodic boundaries at the side of the square, they are periodically reinserted on the opposite boundary.

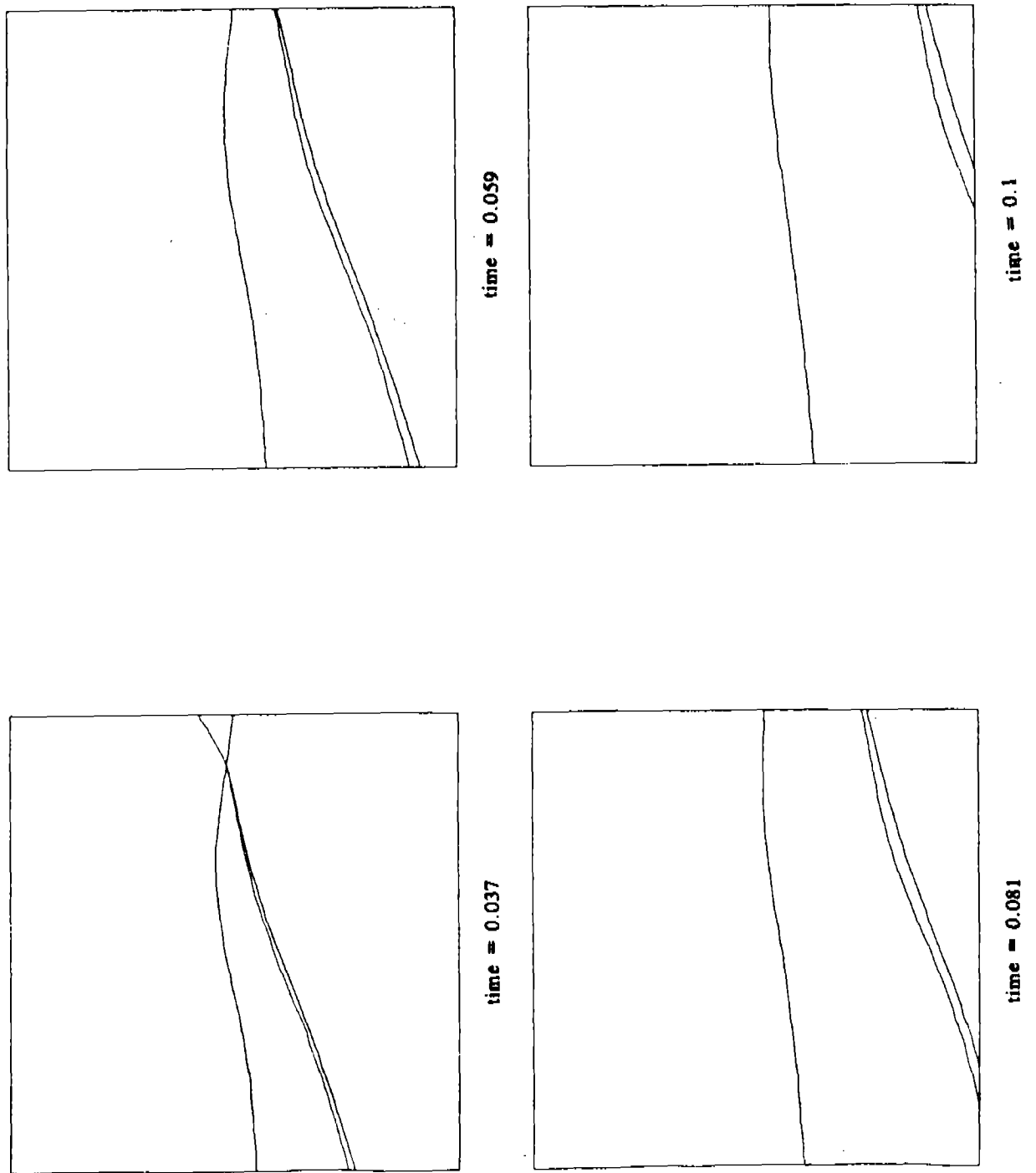


Fig. 6.3
 Propagation of an interior node past a computationally passive boundary on a 20 by 20 grid. Signals from the outside of the computational domain are assumed to be negligible.

BIFURCATION AND STABILITY OF VISCOELASTIC FLUID FLOWS

S. Rosenblat
Department of Mathematics
Illinois Institute of Technology
Chicago, Illinois 60616

ABSTRACT. Results on stability of viscoelastic fluid flows are presented, reviewed and discussed; both linear stability and nonlinear stability (bifurcation) are considered. A central problem in viscoelastic fluid dynamics is that predictions of flow behavior can in general be strongly dependent, both qualitatively and quantitatively, on the choice of the constitutive equation used to relate stress and deformation rate. This is true even for very simple flow configurations, and the acceptability or otherwise of a constitutive law depends on whether its predictions are in accord with experiment. It will be shown in this paper that stability results are also strongly correlated with the particular constitutive equation used, and some implications of this fact are discussed. Detailed results are given for the problem of buoyancy-driven instability in a quiescent horizontal layer.

1. **INTRODUCTION.** The area of hydrodynamic stability can conveniently be subdivided into three categories; these subdivisions are equally relevant to Newtonian and non-Newtonian fluid flows, and apply to both theoretical and experimental approaches.

A. **Linear Stability.** The object is to investigate the stability to infinitesimal disturbances of specific, well-defined basic flow states. In particular, the question of whether a given flow is stable or unstable is usually posed in terms of dimensionless parameter, conveniently denoted R (which may be a Reynolds number, Rayleigh number, Taylor number, etc.), and the aim is to determine a critical value R_c of R which divides stable and unstable flow patterns.

In general, the basic flow is (asymptotically) stable if $R < R_c$ and unstable if $R > R_c$. This component of the area of hydrodynamic stability is thoroughly understood and developed for Newtonian fluid flows.

B. Bifurcation. The object is to investigate the weakly nonlinear evolution into a new state that is associated with the destabilization of a basic state. The theory for this aspect is well developed and applies in a small neighborhood of the critical value of R , that is, for $|R - R_c|$ small. Theorems are available that can predict

whether the new bifurcation state is supercritical (that is, exists for $R > R_c$), subcritical (exists for $R < R_c$) or

transcritical (exists for both $R < R_c$ and $R > R_c$), and

whether it is stable or unstable. The new state may be time-independent or time-periodic, depending on the characteristics of the linear stability problem, and has a spatial structure different from, and more complex than, the basic state.

C. Transition. Most fluid flows experience transition to turbulence when the characteristic parameter R becomes sufficiently large. The problem of transition is currently of great interest, and some understanding of the process has been achieved in recent years. A number of ideas are being explored in this strongly nonlinear regime, in particular the notion that there is a pre-turbulent, chaotic state that appears when $R \gg R_c$. There are several model equations

and a body of theory in dynamical systems that support this idea, but the link with real fluid behavior is only partially established.

2. STABILITY OF VISCOELASTIC FLUIDS. It is convenient to characterize the elasticity of a non-Newtonian fluid by a parameter λ (which may denote a Weissenberg number or a Deborah number); for the purposes of this discussion, $\lambda = 0$ represents a Newtonian fluid and $\lambda > 0$ a viscoelastic fluid. This is a gross simplification; it is rarely the case in practice that a non-Newtonian fluid can be characterized by a single parameter.

As regards stability, it is desirable to summarize the most important issues that arise in relation to viscoelastic fluids by comparison with their Newtonian counterparts.

A. Linear Stability. There are two issues, quantitative and qualitative. The former relates to the question of how the critical value of R_c is changed when $\lambda \neq 0$

in other words, whether the presence of elasticity tends to stabilize or destabilize a given flow field, or to have no effect. This problem has been studied by a great many authors in different flow configurations (see for example

the survey by Pearson [1]). The other, perhaps more important, question is whether new instabilities can appear when $\lambda \neq 0$ which are completely absent when $\lambda = 0$. It is of special significance to know whether such instabilities can arise for $R < R_c$. If they do, it may be associated with a release of elastic energy stored in the basic flow to feed the disturbances. More recently it has been suggested by Joseph and co-workers [2] that such an effect can be due to the change of type from elliptic to hyperbolic of the governing differential equations.

B. Bifurcation. The main issue here is qualitative, whether the presence of elasticity alters the nature of the bifurcating solution. In particular it would be of interest to know whether a solution that is supercritical and stable in the Newtonian case can become subcritical and unstable in the corresponding non-Newtonian case, or conversely. An example of this will be discussed in detail below. Another interesting question is whether the geometry in space of a bifurcating solution is changed due to viscoelasticity.

C. Transition. Although most practical viscoelastic fluid flows operates at relatively low Reynolds numbers, it would nevertheless be important to know how the transition to turbulence occurs in such fluids. This issue has hardly been addressed at all and the answers are unknown.

3. CONSTITUTIVE RELATIONS. The greatest obstacle to understanding the behavior of viscoelastic fluid flows is lack of certainty regarding constitutive relations. For the purposes of solving problems one is obliged to use relatively simple relations rather than general (e.g. functional) laws, but then one encounters the difficulty that a suitable relation for a particular flow of a fluid may not be suitable for a different flow of the same fluid. In addition, of course, different fluids generally require different constitutive laws. In the study of viscoelastic fluid stability there is already evidence that predictions, including qualitative ones, may be relation-dependent, and further examples of this will be given below. This, however, should be regarded as a positive rather than a negative feature, since when coupled with experiments the theory can be used as a yardstick for the validity or otherwise of a particular constitutive equation for certain classes of flows. A good example is the work of Craik [3] who showed that the rest state of a second-order fluid was unstable, which implies that the second-order model should not be used for unsteady flows.

4. STABILITY OF SHEAR FLOWS. There has been much work (almost exclusively linear theory) on the stability of various kinds of rectilinear and circular shearing flows,

and several surveys are available, including that by Pearson [1] and Petrie and Denn [4]. A good deal of this work has been motivated by practical problems relating to flows of polymer solutions and polymer melts.

A. Circular Couette Flow. This is the flow between concentric rotating cylinders, a classical problem in Newtonian fluid stability, and of relevance to the process of screw extrusion. There have been several calculations of linear stability for various constitutive relations [5-7], and the overall conclusion is that for any physically acceptable constitutive model the effect of viscoelasticity is stabilizing; in other words the critical parameter (the Taylor number) has a higher value than in the Newtonian case. The onset of instability is through the exchange of stabilities in these analyses. Under certain conditions, however, [8-9] it appears that overstability is possible. There has been some work on nonlinear instability for this problem [10], but only for very special constitutive equations.

B. Plane Couette Flow. This is the simplest viscometric flow, generated by differential motion of parallel planes. In the case of a Newtonian fluid the flow is stable for all Reynolds numbers to infinitesimal disturbances. There have been many studies of the stability of this flow for various viscoelastic models, with the aim of determining an instability at low Reynolds number but dependent on Weissenberg number. Although there are some positive results [11], it is not at all clear that there is a viscoelastic instability in this flow for a meaningful constitutive relation [12].

C. Plane Poiseuille Flow. This is the flow in a channel under a constant pressure gradient. A number of studies of the stability of this flow [13-18] for various constitutive models have indicated that in general the critical Reynolds number is decreased by the presence of elasticity: a destabilizing effect. The degree of the effect depends on the model. There have also been attempts to identify the presence of a new instability at low Reynolds numbers [18-22], with a view to elucidating the important phenomenon of melt fracture. Although such instabilities are predicted by theory, the results are for special constitutive models and do not in general agree with any experimental results, leaving the question of such instabilities open. Some work has been done on weakly nonlinear theory [23-24], but the conclusions do not appear to be particularly significant.

5. THERMAL CONVECTION. To demonstrate some of the issues raised earlier, we shall discuss in detail the

thermal convection (Benard) problem for a viscoelastic liquid. This is the problem of a horizontal liquid layer heated from below; density varies linearly with temperature and buoyancy drives convection. In the basic state there is no fluid motion and a linear temperature distribution across the layer. The governing equations are Navier-Stokes, energy, continuity and a constitutive relation. The liquid is of infinite horizontal extent, and the parallel horizontal boundaries are taken to be isothermal, non-deformable and stress-free.

The equations (dimensionless) for arbitrary disturbances to the basic state are

$$\text{Pr}^{-1}(\underline{v}_t + \underline{v} \cdot \nabla \underline{v}) = -\nabla p + \nabla \cdot \underline{\tau} + R\theta \underline{\hat{z}} \quad (1)$$

$$\nabla \cdot \underline{v} = 0 \quad (2)$$

$$\theta_t + \underline{v} \cdot \nabla \theta = \nabla^2 \theta + w \quad (3)$$

where $\underline{v} = (u, v, w)$ is velocity, θ is temperature, p is pressure, $\underline{\tau}$ is extra stress, $\underline{\hat{z}}$ is unit vector in the

vertical direction, and R , Pr are respectively the Rayleigh number and Prandtl number defined in the usual way, with the zero-shear-rate viscosity as the reference value. The boundary conditions are

$$\theta = u_z = v_z = w = 0 \quad \text{on } z = 0, 1. \quad (4)$$

To investigate the consequences of various models, we use a hybrid constitutive relation which includes several familiar relations as special cases. We write

$$\underline{\tau} + \lambda \left[\kappa \text{tr} \underline{\tau} + D \underline{\tau} - \frac{1}{2} a (\underline{\tau} \cdot \dot{\underline{\gamma}} + \dot{\underline{\gamma}} \cdot \underline{\tau}) \right] = \dot{\underline{\gamma}} + \epsilon \lambda \left[D \dot{\underline{\gamma}} - \dot{\underline{\gamma}} \cdot \dot{\underline{\gamma}} \right] \quad (5)$$

where $\dot{\underline{\gamma}} = \nabla \underline{v} + \nabla \underline{v}^T$ is the rate of strain tensor, and D is the Jaumann derivative defined by

$$D = \frac{\partial}{\partial t} + \underline{v} \cdot \nabla + \frac{1}{2} (\underline{\omega} \cdot - - \cdot \underline{\omega})$$

where $\underline{\omega} = \nabla \underline{v} - \nabla \underline{v}^T$ is the vorticity tensor. Equation (5) has

four parameters: λ is the Deborah number, a measure of relaxation time; ε ($0 \leq \varepsilon < 1$) is related to the retardation time; a ($0 \leq a \leq 1$) delineates corotational and codeformational models; κ (> 0) is a parameter obtained from network theory arguments. When $\kappa = 0$, equation (5) is an Oldroyd-type model, specifically, a Maxwell model when $\varepsilon = 0$ and a Jeffreys model when $\varepsilon \neq 0$. The model is corotational when $a = 0$ and upper convected when $a = 1$. When $\kappa \neq 0$ and $\varepsilon = 0$ equation (5) becomes the Phan Thien-Tanner model [25].

We now consider the various stability aspects of this problem.

A. Linear Stability. This problem was fully resolved by Sokolov and Tanner [26]. When the nonlinear terms in equations (1), (3) and (5) are deleted, and the time dependence is taken to be proportional to $\exp(\sigma t)$, the problem reduces to an eigenvalue problem for the growth rate σ of disturbances. It should be noted that in the linear approximation the parameters a and κ are irrelevant; only the elasticity parameters enter the problem. Loss of stability occurs when $\sigma = 0$ (exchange of stabilities) or when $\text{Re } \sigma = 0$ (overstability).

The results can be summarized as follows. In the case of exchange of stabilities ($\sigma = 0$) the Rayleigh number for marginal stability is

$$R^{(S)} = (\pi^2 + \alpha^2)^3 / \alpha^2 \quad (6)$$

where α is the wave number of the disturbance. The critical value is $R^{(S)} = 27\pi^4/4$ at $\alpha_c = \pi^2/2$. This is exactly the result for the Newtonian problem, which means that there is an instability that is independent of elasticity. In the case of overstability ($\sigma = i\omega$) there is marginal stability when

$$R^{(P)} = R^{(S)} - c\omega^2 \left[(c\lambda + 1)Pr^{-1} + \varepsilon\lambda c \right] / \alpha^2 \quad (7)$$

with

$$\omega^2 = \frac{c\lambda(1-\varepsilon) - 1 - Pr^{-1}}{\lambda^2(Pr^{-1} + \varepsilon)} \quad (8)$$

where $c = \pi^2 + \alpha^2$. This result depends on the elasticity parameter (as well as on Prandtl number), and, as (8) shows, can only occur if

$$\lambda(1-\varepsilon) > (1+Pr^{-1})/c. \quad (9)$$

In other words, this is a new instability occurring only at sufficiently high elasticity. It is obvious from (7) that $R(P) < R(S)$ at any fixed wave number at which overstability takes place; this means that the first onset of instability may be periodic rather than steady. A graph of $R(S)$ and $R(P)$ as a function of wave number is shown in Figure 1. The solid curve represents $R(S)$, the broken curves $R(P)$ for various parameter values.

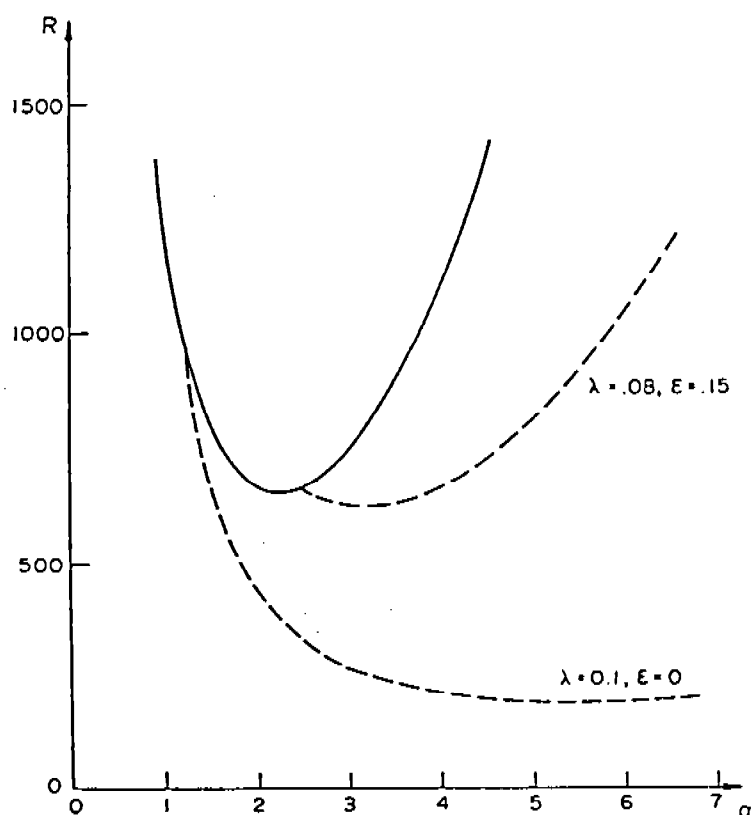


FIGURE 1.

Neutral stability curves for steady onset (solid lines) and periodic onset (broken lines).

B. Bifurcation. A partial study of this problem was performed by Eltayeb [27]. Looking only for two-dimensional weakly nonlinear solutions we can introduce a stream function ψ such that

$$\underline{v} = (\psi_z, 0, -\psi_x) . \quad (10)$$

The governing equations (1) - (3) reduce to

$$Pr^{-1} \left\{ \nabla^2 \psi_t + J(\nabla^2 \psi, \psi) \right\} + R\theta_x - N = 0 \quad (11)$$

$$\theta_t + J(\theta, \psi) + \psi_x - \nabla^2 \theta = 0 \quad (12)$$

where J is the Jacobian and where N is a combination of stress components, namely

$$N = \frac{\partial^2}{\partial x \partial z} (\tau_{11} - \tau_{33}) + \left(\frac{\partial^2}{\partial z^2} - \frac{\partial^2}{\partial x^2} \right) \tau_{13} . \quad (13)$$

The first term on the right of (13) is a contribution from normal stress differences while the second relates to shear thinning. Equations (10) - (12) have to be supplemented by the two-dimensional forms of the nonlinear constitutive relations (5).

The problem is solved by perturbation in terms of a small bifurcation parameter μ . We write

$$R - R_c = \mu^2 R_2 \quad (14)$$

where R_c is the critical value for either steady or periodic onset. A solution that exists for $R_2 > 0$ is called

supercritical and is called subcritical if it exists for

$R_2 < 0$. All field quantities are expanded in powers of μ , for example

$$\theta = \mu \theta_1 + \mu^2 \theta_2 + \mu^3 \theta_3 + \dots \quad (15)$$

The quantity θ_1 has the form

$$\theta = A \cos \alpha x \sin \pi z \quad (16)$$

where A is the amplitude, to be determined. A standard perturbation procedure leads to a solvability condition at order μ^3 .

We summarize the results for the case of steady onset of convection. Details and results for the case of periodic onset can be found in [28]. The solvability condition leads to an equation of the form

$$R_2 A - K A^3 = 0 \quad (17)$$

where K is a constant that depends on all parameters of the problem. The bifurcating solution will be supercritical and stable if $K > 0$, and will be subcritical and unstable if $K < 0$. In the Newtonian case it is known that $K = p > 0$, and the solution is supercritical and stable. In the present problem we compute an expression for K , namely

$$K = p + \lambda^2 (1-\epsilon) \left[m - (a^2 - 2a\kappa)n \right] \quad (18)$$

where m, n are positive numerical constants. It follows from (18) that there are values of the viscoelastic parameters which allow subcritical bifurcation. We find, for example, that if $a > 0.75$ (close to the upper convected model), $\kappa = 0$ and $\lambda > 0.03$ with $\epsilon = 0$, then subcritical bifurcation occurs. Other combinations give the same result, but if a is sufficiently small or κ sufficiently large then only supercritical bifurcation can occur.

Results such as these may be a partial test of the viability of a proposed constitutive relation for a particular liquid. Thus Liang and Acrivos [29] studied onset of convection in a polyacrylamide solution and found only supercritical bifurcation. This observation, coupled with the formula (18), would exclude certain models, or certain parameter ranges, for this liquid.

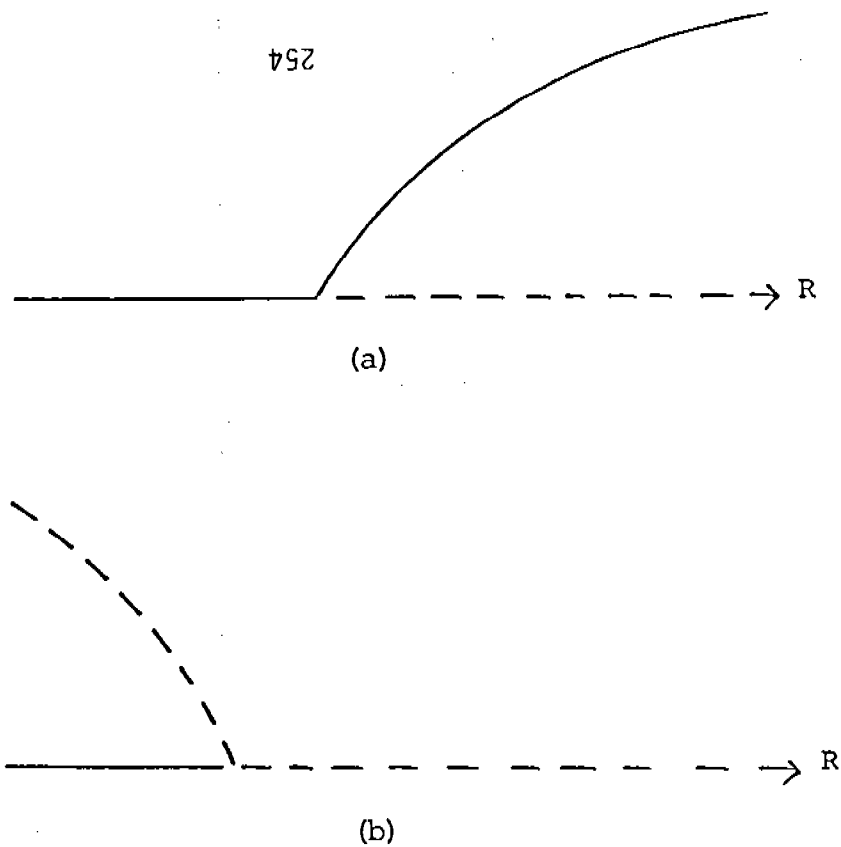


FIGURE 2

Schematic bifurcation diagram for (a) supercritical bifurcation and (b) subcritical bifurcation. Solid lines represent stable solutions and broken lines represent unstable solutions.

C. Transition. A formal study of transition has not yet been achieved, even in Newtonian fluids. In recent years, however, there has been great interest in a severely truncated Fourier series representation of the field quantities that leads [30] to a system of three nonlinear ordinary differential equations having interesting properties. These equations, known as the Lorenz system, exhibit aperiodic (chaotic) solutions that may be a valid behavioral model of how transition takes place. It is of interest to derive the analogous system for the viscoelastic problem, and to study its properties.

To do this we set

$$\theta = A_1 \cos \alpha x \sin \pi z + A_2 \sin \alpha x \sin \pi z, \quad \psi = B_1 \sin \alpha x \sin \pi z$$

$$N = M_1 \sin \alpha x \sin \pi z \quad (19)$$

in the equations of motion and the constitutive relations, and then truncate to retain only those Fourier components exhibited in (19). The quantities A_1 , A_2 , B_1 , M_1 , are time-dependent amplitudes. The procedure leads to a system of four ordinary differential equations, namely,

$$\begin{aligned} \text{Pr}^{-1} c B_1' + \alpha R A_1 + M_1 &= 0 \\ A_1' + c A_1 + \alpha B_1 + \pi \alpha A_2 B_1 &= 0 \\ A_2' + 4\pi^2 A_2 - \frac{1}{2} \pi \alpha A_1 B_1 &= 0 \\ M_1 + \lambda M_1' - c^2 (b_1 + \epsilon \lambda B_1') &= 0 \end{aligned} \quad (20)$$

It is noteworthy that of the four elasticity parameters in the original system, only two (λ and ϵ) appear in (20); this is due to the form of the truncation. Also, when $\lambda = 0$ the above reduces precisely to the three equations of the Lorenz system.

The linear and weakly nonlinear stability analysis of the null solution of (20) gives the same results as obtained

for the full system above. When $\lambda(1 - \epsilon) < (1 + \text{Pr}^{-1})/c$ there is a steady bifurcation from $R = R_c$ and in this case

the solution is supercritical. When $\lambda(1 - \epsilon) > (1 + \text{Pr}^{-1})/c$ there is a periodic bifurcation which is also supercritical and unstable. The behavior away from the neighborhood of the bifurcation point has to be determined by numerical integration of (20). The results can be summarized as follows:

(i) When $(1 - \epsilon) < (1 + \text{Pr}^{-1})/c$ the steady bifurcating solution loses stability at a value of R that depends on Pr and the elastic parameters, and a periodic solution bifurcates from it. In the Newtonian case ($\lambda = 0$) this occurs at $R/R_c \approx 24$ when $\text{Pr} = 10$, and the periodic solution

is subcritical. Chaotic solutions are found in regions of parameter space where no stable solutions exist, as shown in

Figure 3. In the viscoelastic case, the steady supercritical solution loses stability at a value of R/R_C

less than in the Newtonian case; in fact this value decreases towards $R/R_C = 1$ as $\lambda(1 - \epsilon) \rightarrow (1 + Pr^{-1})/c$. In

addition the bifurcating periodic solution begins as supercritical and stable, but then turns around, as shown. Chaotic solutions have been computed quite close to the original bifurcation point $R = R_C$ for the appropriate values of the elasticity parameters.

(ii) When $\lambda(1 - \epsilon) > (1 + Pr^{-1})/c$ the periodic solution emerging from the null solution is supercritical and stable, and remains so; it does not lose stability. Moreover in this case there appears to be no chaotic solutions at all, as far as we can determine.

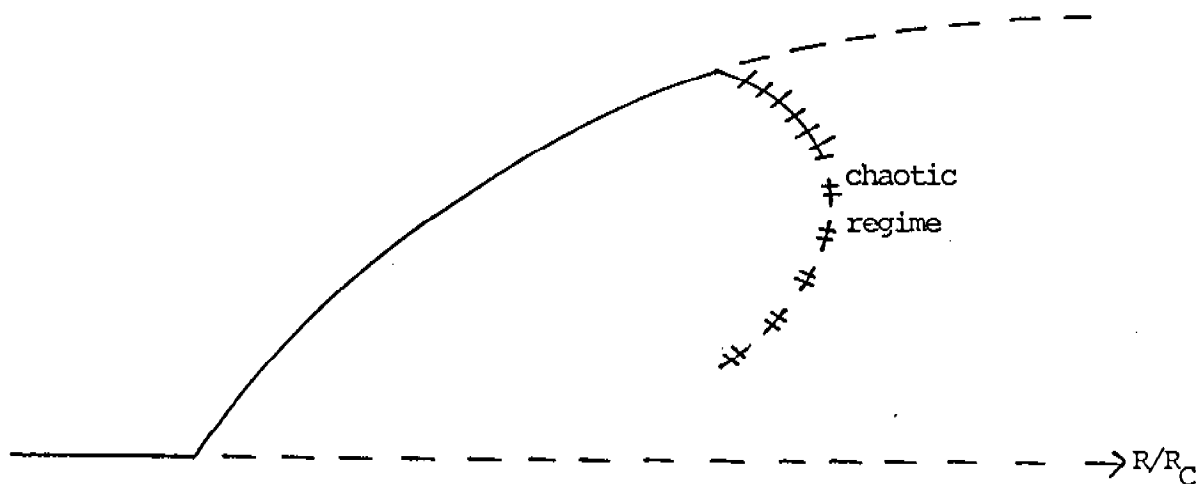


FIGURE 3

Schematic bifurcation diagram for equations (20). Stable steady solutions are denoted by ———, unstable steady solutions by - - -, stable periodic solutions by + + + +, unstable periodic solutions by x x x x.

Thus elasticity acts in two different ways in the system (20): it first advances the position where chaotic solutions appear and then, for higher values, it seems to suppress them altogether. This phenomenon warrants further study.

6. SUMMARY. It is evident that both qualitative and quantitative changes can occur in stability problems for viscoelastic flows by comparison with their Newtonian counterparts, but that the results are often strongly dependent on constitutive relations. This fact can be used as a partial test for the validity of a constitutive hypothesis when coupled with appropriate experiments.

In the particular case of the thermal convection problem it is found that a new, periodic instability of the null solution may appear in a sufficiently elastic fluid, and also that the bifurcation into convection may be subcritical for certain constitutive relations. The detailed formulas demonstrated above could be used as measure of fluid behavior and properties.

7. ACKNOWLEDGEMENT. The work described in this paper was supported by the Army Research Office under Contract No. DAAG 29-82.K-0061.

REFERENCES

1. J. R. A. Pearson, Ann, Rev. Fluid Mech., 8, 163, 1976.
2. M. Ahrens, D. D. Joseph, M. Renardy and Y. Renardy, Rheol. Acta. 23, 345, 1984.
3. A. D. Craik, J. Fluid Mech. 33, 33, 1968.
4. C. J. S. Petrie and M. M. Denn, AIChE J., 22, 209, 1976.
5. C. Miller, Univ. of Mich. Ph.D. Diss., 1967.
6. F. J. Lockett and R. S. Rivlin, J. de Mecan., 7, 475, 1968.
7. M. M. Smith and R. S. Rivlin, J. de Mecan., 11, 70, 1972.
8. Z. S. Sun, Univ. of Del. Ph.D. Diss., 1972.
9. H. Giesekeus, Prog. in Heat Mass Transfer, 5, 187, 1972.
10. M. M. Denn, Z. S. Sun and B. D. Rushton, Trans. Soc. Rheol. 15, 415, 1971.
11. U. Akbay and S. Sponagel, Rheol. Acta, 20, 579, 1981.
12. M. Renardy, 3rd Army Conf. on Applied Math. and Comp., Atlanta, 1985.
13. G. Tlapa and B. Bernstein, Phys. Fluids, 13, 565, 1970.
14. K. C. Porteous and M. M. Denn, Trans. Soc. Rheol., 16, 295, 1972.
15. Chan Man Fong, Rheol. Acta, 7, 324, 1968.
16. D. H. Chun and W. H. Schwarz, Phys. Fluids, 11, 5, 1968.

17. J. Platten and R. S. Schechter, *Phys. Fluids*, 13, 832, 1970.
18. J. R. A. Pearson and C. J. S. Petrie, *Proc. Fourth Int. Cong. Rheol.*, 3, 265, 1965.
19. L. V. McIntire, *J. Appl. Polymer Sci.*, 16, 290, 1972.
20. W. S. Bonnett and L. V. McIntire, *AIChE J.*, 21, 901, 1975.
21. R. Rothenberger, D. H. McCoy and M. M. Denn, *Trans. Soc. Rheol.*, 17, 259, 1973.
22. T. C. Ho and M. M. Denn, *J. Non-Newt. Fluid Mech.*, 3, 179, 1977.
23. K. C. Porteous and M. M. Denn, *Trans. Soc. Rheol.*, 16, 309, 1972.
24. L. V. McIntire and C. H. Lin, *J. Fluid Mech.*, 52 273, 1972.
25. N. Phan Thien and R. I. Tanner, *J. Non-Newt. Fluid Mech.*, 2, 353, 1977.
26. M. Sokolov and R. I. Tanner, *Phys. Fluids* 15, 534, 1972.
27. I. A. Eltayeb, *Proc. Roy. Soc. Lond. A*, 356, 161, 1977.
28. S. Rosenblat, to appear, 1985.
29. S. F. Liang and A. Acrivos, *Rheol. Acta*, 9, 447, 1970.
30. E. N. Lorenz, *J. Atmos. Sci.*, 20, 130, 1963.

CRACK SOLUTIONS AND DUCTILE FRACTURE CRITERIA

Dennis M. Tracey and Colin E. Freese

Mechanics of Materials Branch
Army Materials and Mechanics Research Center
Watertown, Massachusetts 02172-0001

ABSTRACT. Numerical solutions for a group of elastic-plastic crack problems are discussed. The problems consider blunted flaws in infinite plates under remote monotonically increasing tension. The solutions were obtained using a combined finite element-analytic stress function formulation. Results related to critical stress and energy release rate theories of ductile fracture are discussed.

I. INTRODUCTION. The results presented in this report supplement other results presented by the authors on the topic of elastic-plastic stress states near the ends of blunted, cracklike flaws, Ref.(1,2). The inclusion of crack tip geometry in analytical studies is a complication which has been largely avoided in the mechanics of fracture literature. Noteworthy exceptions are the papers of Rice and Johnson(3) and McMeeking(4) which have considered the finite deformation effects at the tip of a sharp crack. The goal of our work has been to determine the important aspects of the elastic-plastic crack solution related to the presence of the traction free crack tip surface so that guidelines to the applicability of sharp crack solutions to ductile fracture problems can be established.

Our work has concentrated on the character of the near tip elastic-plastic stress and strain fields and how these compare with the results of sharp cracks. Specific models treated have been flaws with semicircular and circular tips (U and keyhole) and the very long and narrow elliptical flaw.

In Ref.(2) we reported the intriguing result that the maximum stress of the logarithmic spiral slipline solution is not achieved for either the U or keyhole tip due to a limited range of plasticity along the tip surface. This is an important finding with implications related to critical stress theories of fracture such as discussed by Ritchie, Knott and Rice(5). Here we provide results indicating the range of applicability of the fully plastic slipline solution and the character of the elastic-plastic stress distribution for the cracklike ellipse.

Another important aspect of the solutions which relate to energy theories of fracture is the variation of the J integral near the crack tip, Ref.(6). J is a path integral defined in terms of the strain energy density $W(\underline{\epsilon})$, displacement gradient and stress traction along an arbitrary path L which starts at a point on the lower crack flank and ends at a point on the upper crack flank. With the crack length in the x direction, s arclength increasing counterclockwise, \underline{n} outward pointing unit normal, \underline{T} the traction vector and \underline{u} the displacement vector,

$$J = \int_L (Wn_x - \underline{T} \cdot \partial \underline{u} / \partial x) ds$$

J was introduced by Rice(7) who demonstrated that for constitutive theories which have stress derivable from W, J is path independent and equal to the rate of decrease of potential energy with respect to crack length. As such, J can be expected to characterize the severity of the crack tip deformation field,

while incipient crack extension might be experimentally correlated and predicted in application in terms of a material dependent critical value of J .

A strain energy density function does not exist for materials which yield and flow according to the Prandtl-Reuss theory. This is the theory most commonly employed in studies of metal deformation and it was used in this work. However, if throughout the domain proportional stressing results, the solution corresponds to the nonlinear elastic solution of the problem. Of course a strain energy function does exist for nonlinear elasticity. This fosters interest in evaluating J with W defined as the stress working density

$$\int_0^{\epsilon} \underline{\sigma} \cdot d\underline{\epsilon}$$

The interest lies in establishing the degree of path dependency of J for the elastic-plastic case and determining how the crack tip value J_{tip} differs from the value calculated over paths within the elastically deformed material away from the tip. J computations have been performed for the U-tip flaw and results are discussed below.

II. NUMERICAL FORMULATION. The flaw of our study is taken to be crackline with a length $2a$ greatly exceeding the root radius of curvature ρ . The flaw is isolated within an infinite metallic sheet which is under plane strain constraint with a tensile load applied far from the flaw. The metal deforms according to the Prandtl-Reuss equations which have linear elastic response within the Mises yield surface and plastic flow following the normality rule. These constitutive relations are incremental in form which means that the solution to our nonlinear problem must be obtained by progressing stepwise along the load path. The nonhardening model is used, so that plastic flow occurs under a constant value of equivalent stress equal to the material's uniaxial yield stress Y . Besides the yield stress, the only other material

properties entering the analysis are Young's modulus E and Poisson's ratio ν , and the latter was taken equal to 0.3. For general applicability, the results are presented in dimensionless form in terms of Y and E .

The analysis was conducted using a numerical formulation designed for this problem from aspects of the finite element and stress function boundary collocation methods. In a region surrounding the flaw root, where plastic deformation is anticipated, finite element approximations are made for the displacement field. Over the remainder of the infinite domain, the response is elastic and is represented by an analytic stress function power series approximation. Boundary collocation techniques are used to couple the equations governing in the two regions. At each step of the analysis, discrete unknowns are nodal displacement increments in the finite element region and coefficients of the power series in the elastic region.

The formulation has been discussed in Ref.(1). It follows from the work of Bowie and Freese(2) on mapping-collocation techniques using elastic stress function theory and the work of Tracey and Freese(9) on elastic-plastic finite element analysis. Its appeal lies in the ease by which infinite regions can be accommodated with discretization necessary only at the ends of the flaw. Mapping and analytic continuation are used so that the traction free crack boundary condition is implicitly satisfied. There is no need for discretization and collocation along the crack surface in the elastic region.

The standard mapping function which transforms the ellipse with semi-axes a, b to the unit circle $|\zeta|=1$ in a ζ parameter plane was used in the elastic formulation. The U-tip and keyhole problems were treated by considering the flaws as slits ($b=0$) in the elastic region. The finite element mesh served to define the crack tip shape in the problems. In the U-tip problem, within the finite element region the slit opens to a parallel faced,

semi-circular ended slot of length $2a_0$; while in the keyhole problem, a split circular boundary was connected to the ends of the slit. While a more natural U-tip model would have a uniform opening along the entire crack length, the mapping function which would transform this flaw onto the unit circle is not known. Nonetheless, there is little reason to expect that this case would have a solution very much different than the case considered.

Whereas the general planar elasticity problem requires determination of two analytic stress functions, continuation reduces the problem to finding a single function $\phi(\zeta)$ which satisfies the remote stress condition and the equilibrium and compatibility conditions along the finite element interface. While formally there is an interface encircling each end of the flaw, symmetry allows treatment of a single quadrant of the plane.

The interface is defined in the auxiliary plane as a circle of radius R centered at the ends of the slit, so that the problem in the elastic region is to find ϕ outside the disks $|\zeta \pm 1| < R$. The interface maps onto a smooth non-circular contour ending on the faces of the ellipse (slit). For the elliptical flaw R was chosen to have the interface a distance 125ρ from the tip at 70° from the length direction. The circular tip problem had R chosen so that the interface was at a distance of 75ρ at 70° .

The nature of formulation is such that the computational task increases with the value of R in that a larger region must be discretized when R is increased while a higher load level can then be accommodated. The maximum load that can be incrementally reached is determined in the course of the analysis according to when the plastic zone extends to the interface.

The power series approximation for $\phi(\zeta)$ that was chosen consists of two parts. The first part corresponds to the known solution for an elliptical flaw in a purely elastic tension field. The second part serves to represent the perturbation from this elastic solution near the root. In terms of the remote stress increment ΔT and undetermined coefficients α_n , the approximation has the form

$$\phi(\zeta) = \Delta T \{ (a + b)\zeta - (3a + b)/\zeta \} / 8 + \sum_{n=1}^m \alpha_n \zeta / (\zeta^2 - 1)^n$$

The first term is expected to adequately represent the solution far from the plastic zone. The additional series is expanded from the flaw ends $\zeta = \pm 1$ and its terms vanish at infinity, consistent with its role of representing the local deviation from the exact elastic solution. There are m coefficients α_n , and they are real numbers due to the conditions of symmetry across the x and y axes. Sufficient accuracy was found in the analyses reported here by using 20 terms in the series approximation.

Reasonably fine element grids were used in the analyses. A mesh consisting of approximately 1200 nodes and 4700 triangular elements was used in the elliptical flaw problem, while for the other problems the mesh consisted of approximately 1000 nodes and 2000 elements. Element dimensions in each case increased from the tip and at the tip edge lengths were less than $\rho/10$.

The usual stiffness approach employing the principle of virtual work was used to assemble the governing equations in the finite element region. The load vector in this system is made up exclusively of force increments acting on interface nodes. These represent the load transfer across the interface. A partial Gaussian elimination provides a reduced linear system of

equations relating these interface force increments to the interface nodal displacement increments. The force increments are expressed in terms of α_n using the elasticity theory stress equations and the conventional finite element consistent load procedure. Likewise, the displacement increments are expressed in terms of α_n . A series with fewer terms than finite element interface degrees of freedom is chosen so that the system is over-determined and solution is by the method of least squares. The finite element nodal displacement increments are calculated once α_n are computed and then strain and stress increments are computed throughout the finite element mesh.

Details of the finite element formulation have been thoroughly discussed in Ref.(9). An average stiffness approach is employed to accommodate plastic zone and flow rule changes during a step. A nonlinear problem is posed at each step since the averages are defined in terms of the undetermined nodal displacement changes and the solution is found by iteration. We have labeled it an "incremental secant stiffness" formulation to distinguish it from tangent stiffness formulations which define equations on the basis of the current state. Corrective techniques are necessary with the tangent formulations to force the stress solution to satisfy the yield criterion with the result that significant load imbalance errors can ensue. The incremental secant stiffness approach on the other hand guarantees that the yield criterion is satisfied at the end of each step, while load imbalance is controlled by the iteration convergence tolerance.

An adaptive load incrementation algorithm is used to control the load path discretization error. Load step size is treated as a variable and the solution is determined according to a constraint selected to limit the constitutive law changes during a step. For the problems described below the constraint limited the maximum deviatoric stress increment modulus to 0.05Y. In the case of the circular tip flaws, 50 steps were taken to have the

plastic zone extend to within an element of the interface and this corresponded to a load level $T=.28Y$. The elliptical flaw problem required 51 steps to a load $T=.35Y$ for the plastic zone to approach the interface.

III. ELASTIC PLASTIC SOLUTIONS. We discussed in Ref.(2) the finding that for the circular tip flaws the material along the crack tip surface between approximately 67° and 90° elastically unloaded after having yielded. It appears that the expansion of the plastic zone into the region above the crack flanks creates a load shedding mechanism which results in the unloading. Directly related to this limited crack surface yielding behavior is the nature of the stress distribution ahead of the tip. The maximum stress value of $2.97Y$ at $x=3.81\rho$ predicted from slipline theory under the assumption of fully plastic conditions up to the flanks is not realized. Instead it was found that a stress maximum equal to $2.57Y$ develops at $x=2.76\rho$. The logarithmic spiral stress distribution holds only over the range from the crack tip ($x=0$) to $x=2.18\rho$.

Figure(1) illustrates the elastic-plastic limitations to the fully plastic stress solution for the circular tip flaws and also shown are the recently obtained results for the elliptical flaw. The stress values are plotted relative to the yield stress and the distance scale is normalized with respect to the root radius. The solid curves are the slipline predictions assuming that yielding spreads along the crack flanks. The maximum stress of the elastic-plastic solution is indicated for each case as is the point at which the fully plastic solution ceases to apply.

Compared to the circular tip flaws, the elliptical flaw is seen to have a less severe stress state, with a maximum stress equal to $2.44Y$ at a distance of 4.98ρ ahead of the flaw. Figure(2) provides the complete elastic-plastic distribution for the ellipse over a distance 10ρ ahead of the tip. It is seen that the solution departs from the fully plastic curve at $x=4.19\rho$.

The results are plotted from discrete element data along the x-axis at a load level of 0.35Y. Actually, the distribution shows little change over the distance plotted beyond a remote tension value of 0.30Y.

J integral computations were carried out for the U-tip flaw problem to establish the degree of path dependence which follows from the non-proportional stressing in the problem. Paths L were chosen to be piecewise linear contours along element edges which encircle the crack tip, as illustrated in Figures(3,4). The problem of data retrieval was managed by a scheme which employed element face numbers defined by combining the numbers of the nodes on each face of the mesh. Paths were specified by listing the nodes along each path so that elements contributing to each arc of the path were identified by a face number test. Stress, strain and displacement data from the two constant state triangles sharing each segment were averaged and this provided the data needed to compute the edge contributions to J.

Results are given in Figure(5) corresponding to the 18 paths drawn in Figure(4) which are from the crack tip surface ($r=\rho$) to $r=16.8\rho$. The results are plotted with J normalized by a reference value J_{ref} which was taken to be the value for a perfectly sharp crack (slit) with length $2a$ and remote tensile stress T ,

$$J_{ref} = (1 - \nu^2) \pi a T^2 / E$$

The numerical error of the finite element solution and the method of computation of J can be gauged from the path dependency found for the elastic solution. The elastic values of J ranged from 1.007 to .995 J_{ref} with the maximum at $r=1.5\rho$ and only minor variations on paths beyond 5ρ . We might conclude then that the model and computational scheme allow J computations which are

approximately 1% in error.

It is seen that before crack tip unloading, at $T=.16Y$, the value of J_{tip} is 6% below J_{ref} . At the load level $T=.28Y$, the tip value of J is 10% below the reference value. The plot illustrates the trend which has significant J variations over an increasing region from the tip as load level increases.

IV. SUMMARY. The elastic-plastic blunt flaw solutions indicate that the plastic zone extends over a limited portion of the crack surface. This results in a stress distribution ahead of the flaw which is less severe than fully plastic slipline analysis would suggest. The analyses were run to load levels high enough to achieve the asymptotic stress distributions over a 10ρ distance ahead of the flaw. While the circular tip flaws have a maximum stress of $2.57Y$ at 2.76ρ below the crack surface, the elliptical flaw has a maximum stress of $2.44Y$ at 4.9ρ .

J integral computations for the U-tip flaw indicate that there is significant path dependency. The value of J_{tip} in comparison to the remote value of J changes with load level consistent with the evolution of the solution to the fully developed crack tip region stress state. The results suggest that fracture criteria based on J should account for differences in the tip and remote values of J when representing the severity of the deformation state in the fracture prone crack tip material.

REFERENCES

1. Tracey, D. M. and Freese, C. E., (1982), "Cyclic Plasticity Near a Cracklike Elliptical Flaw," Mechanics of Materials, 1, 151-159.
2. Tracey, D. M. and Freese, C. E., (1985), "Importance of Crack Tip Shape in Elastic-Plastic Fracture Analysis," Proc. Second Army Conf. on Applied Math. and Computing, 359-371.
3. Rice, J. R. and Johnson, M. A., (1970), "The Role of Large Crack Tip Geometry Changes in Plane Strain Fracture," in Inelastic Behavior of Solids, M. F. Kanninen et. al. Eds., McGraw-Hill, 641-672.
4. McMeeking, R. M., (1977), "Finite Deformation Analysis of Crack Tip Opening in Elastic-Plastic Materials and Implications for Fracture Initiation," J. Mech. Phys. Solids, 25, 357-391.
5. Ritchie, R. O., Knott, J. F. and Rice, J. R., (1973), "On the Relationship Between Critical Tensile Stress and Fracture Toughness in Mild Steel," J. Mech. Phys. Solids, 21, 395-410.
6. Begley, J. A. and Landes, J. D., (1972), "The J-Integral as a Failure Criterion," in Fracture Toughness, ASTM STP 514, 1-20.
7. Rice, J. R., (1981), "A Path-Independent Integral and the Approximate Analysis of Strain Concentration by Notches and Cracks," J. Appl. Mech., 35, 379-396.
8. Bowie, O. L. and Freese, C. E., (1978), "Analysis of Notches Using Conformal Mapping," in Mechanics of Fracture 5: Stress Analysis of Notch Problems, ed. Sih, G. C., Noordhoff, 69-134.
9. Tracey, D. M. and Freese, C. E., (1981), "Adaptive Load Incrementation in Elastic-Plastic Finite Element Analysis," Computers and Structures, 13, 45-53.

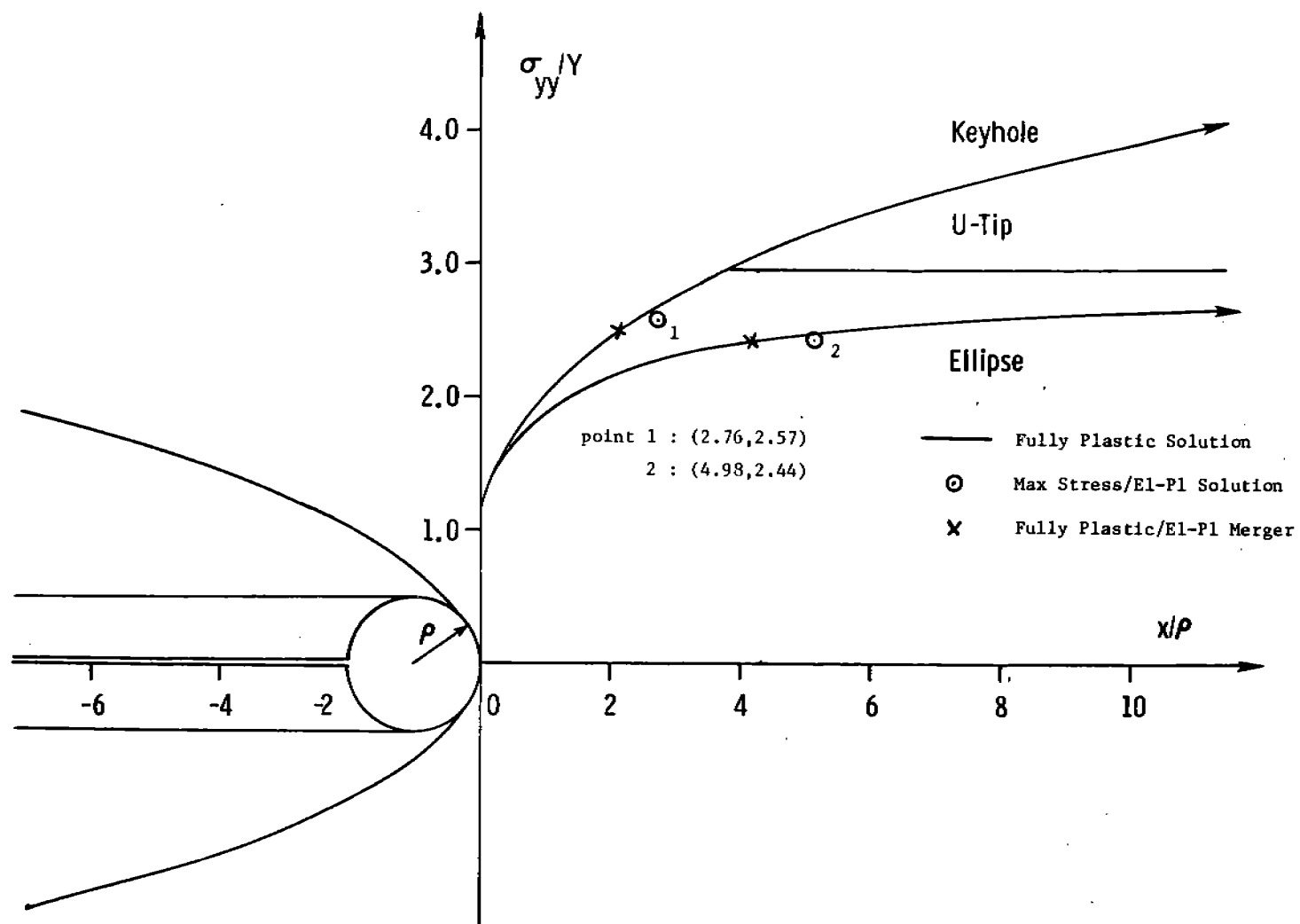


Fig.1 Elastic-Plastic Limitations to Fully Plastic Stress Solutions for Three Crack Models

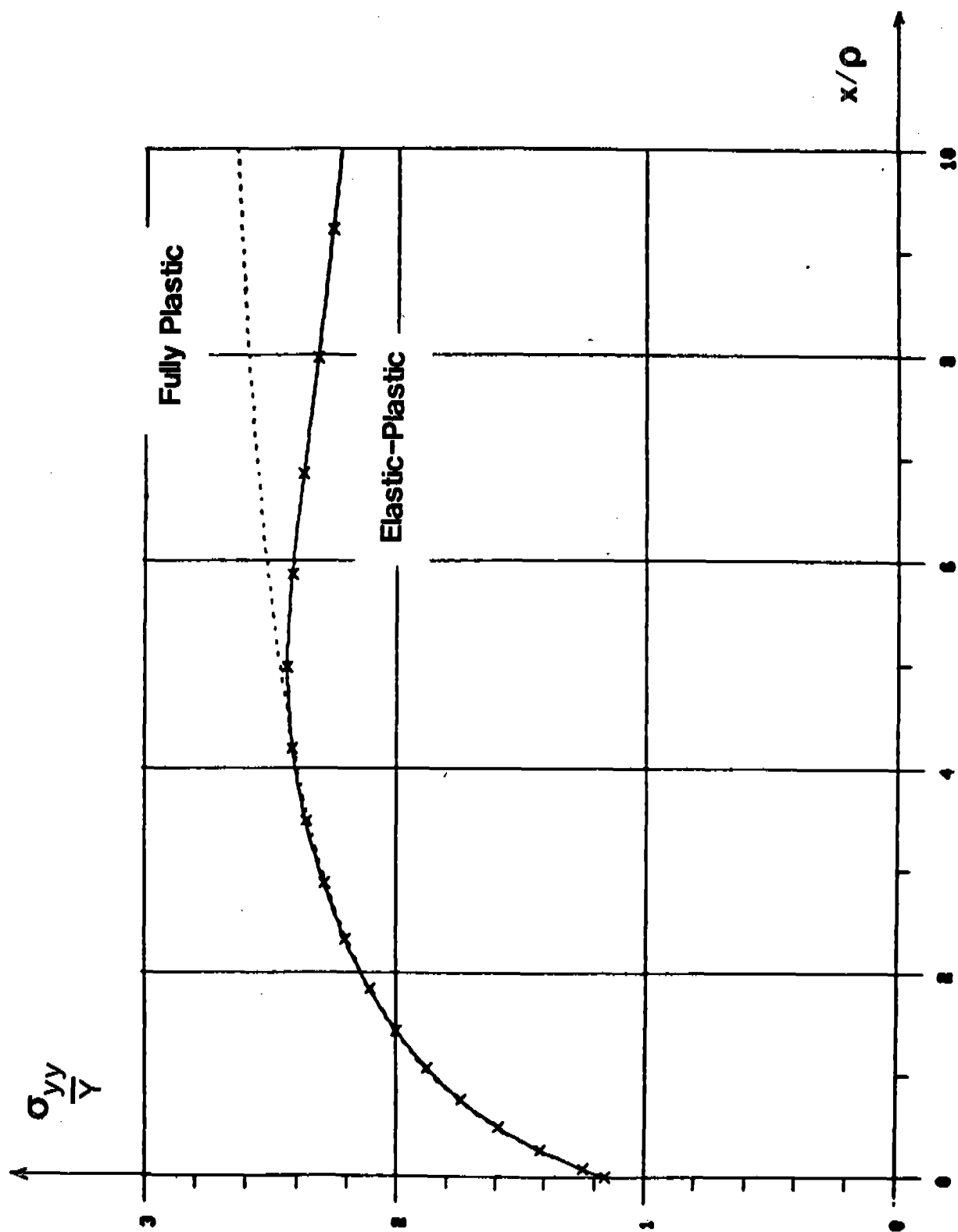


Fig.2 Stress Variation Near Elliptical Flaw

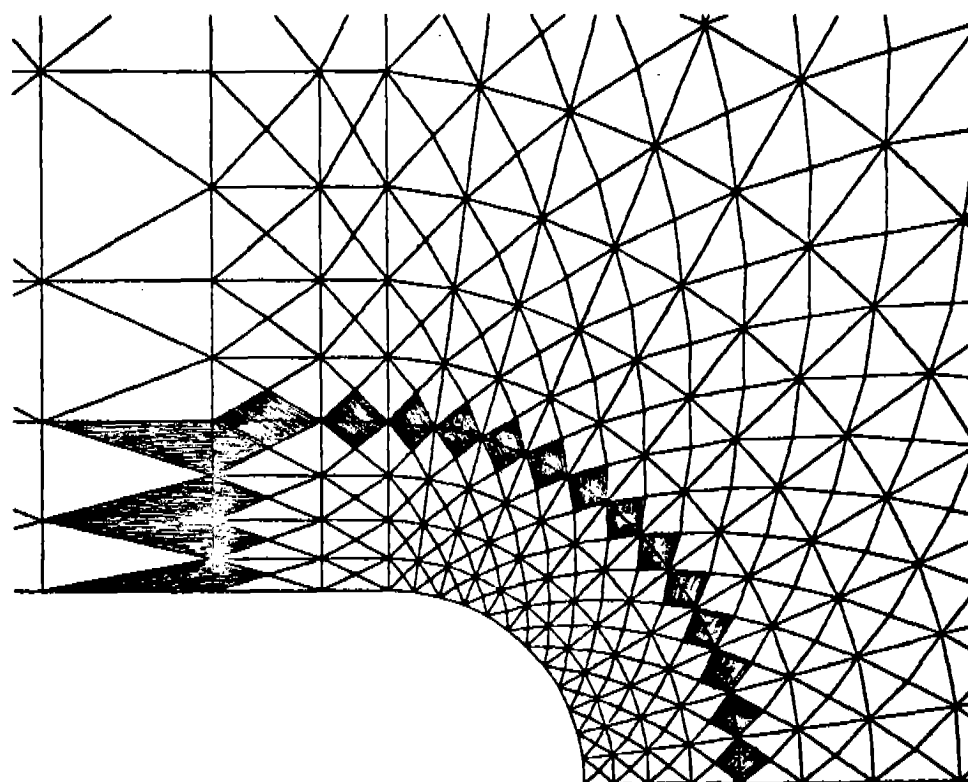
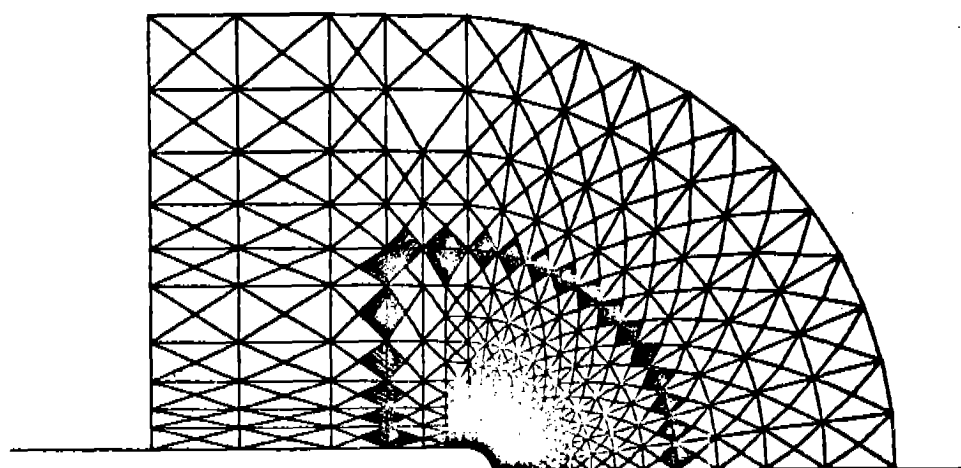


Fig.3 J Integral Contours Along Common Edges of Shaded Triangular Elements

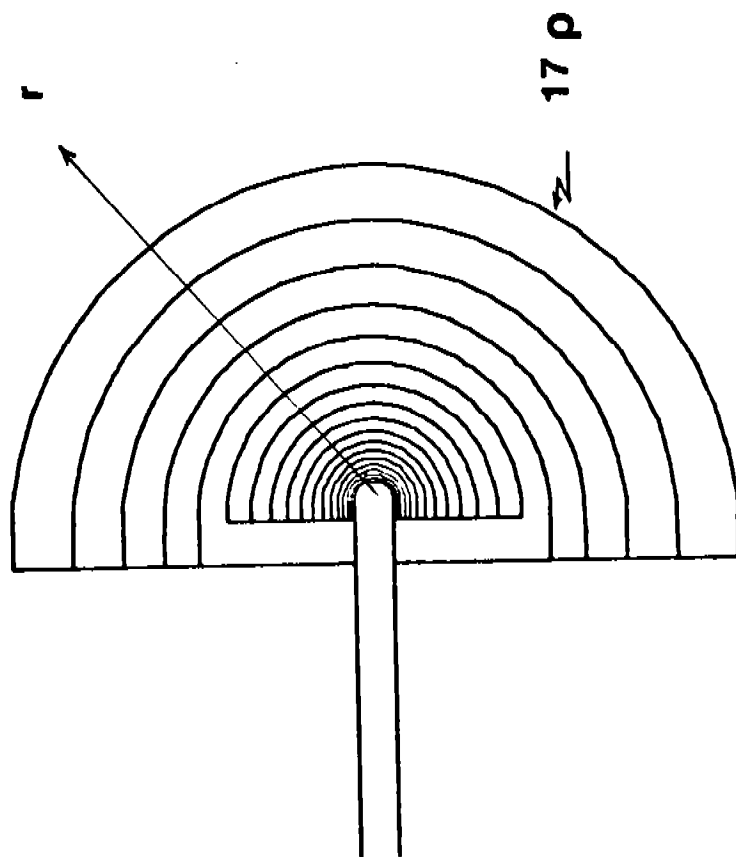


Fig.4 Contours Used in J Integral Computations

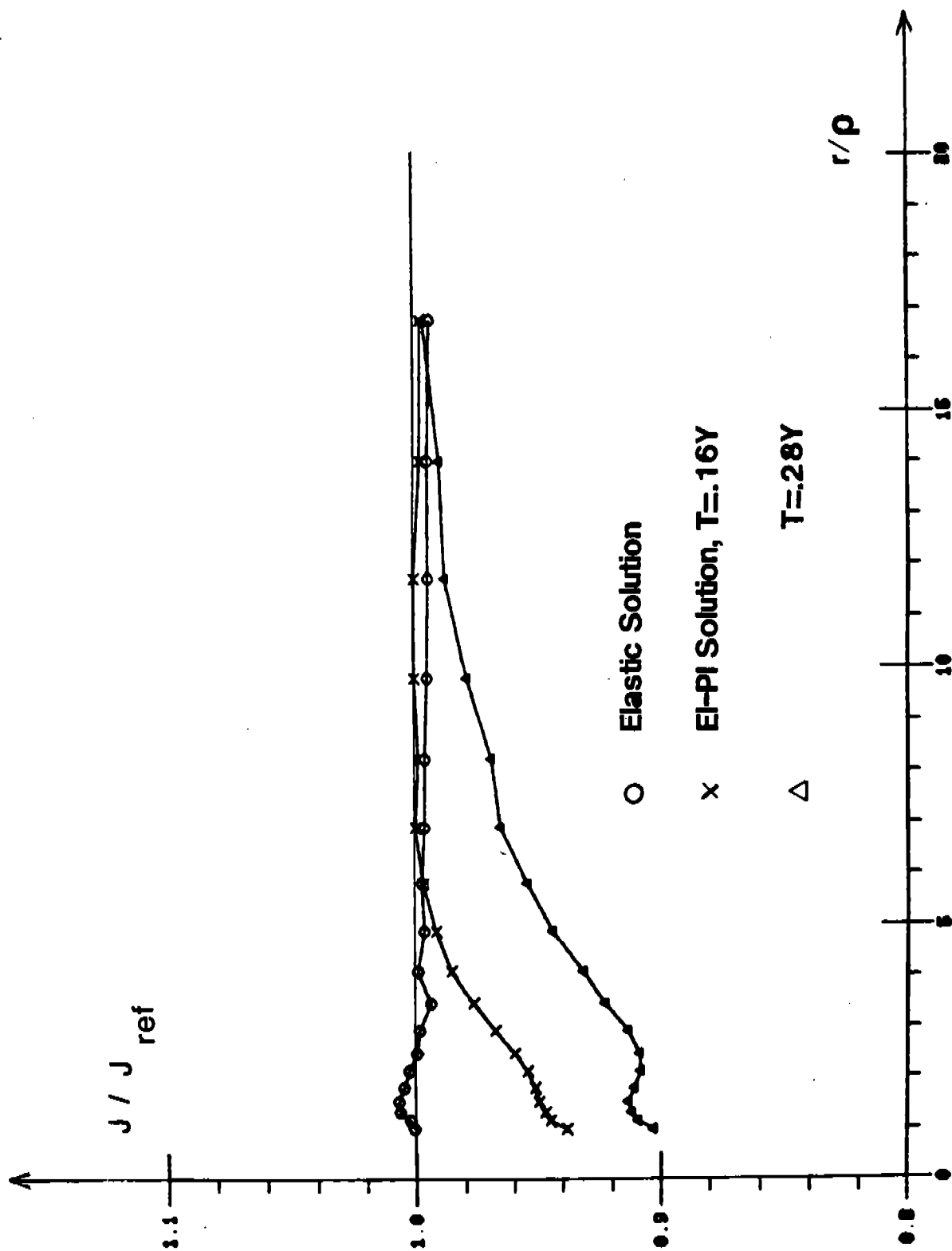


Fig.5 J Path Dependency at Three Load Levels, U-Tip Flaw

THE BAUSCHINGER EFFECT ON STRESS INTENSITY FACTORS
FOR A RADIALLY CRACKED GUN TUBE

S. L. Pu and P. C. T. Chen
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189-5000

ABSTRACT. The theoretical predicted fatigue life of a high-strength steel tube which has undergone an autofrettage procedure is significantly higher than the experimental prediction. To account for the discrepancy, attention is now turned to developing better elastic-plastic models for a high strength steel. An improved material model shows that reverse yielding may occur in the inner portion of the tube. This reverse yielding reduces the residual compressive hoop stress considerably which has an adverse effect on bore crack propagation. This study considers the stress intensity factors due to a radial crack taking the Bauschinger effect into consideration.

The elastic-plastic interfaces during loading and unloading in the autofrettage process divide the tube into three-ring regions. The residual stress distribution in each region is quite different. When a crack grows from one region into another, the previous method using functional stress intensity fails. A new method is used to obtain stress intensity factors for a radial crack growing out of the reverse yielding zone. This approach is based on crack face weight functions obtained by Sha using stiffness derivative finite element techniques coupled with singular crack-tip elements.

I. INTRODUCTION. An early brittle failure of a cannon tube during the 1960's prompted a renewed interest in fracture analysis of cannon pressure vessels [1]. After careful investigation of the cause of fracture, some basic design changes were made to prevent any further failures. One of the design changes was to introduce a compressive residual stress near the inner radius of the cannon by an autofrettage process. Considerable efforts have been made to predict the residual stress distribution in an overstrained tube based on various material models [2-5]. In order to confirm the theoretical predictions of residual stresses, experimental methods [6,7] have been initiated in our laboratory.

Since the current cannon tube design uses a wall ratio close to two, the maximum compressive residual stress at the bore for a fully autofrettaged tube is about 85 percent of the yield stress σ_0 of the material. Most of the earlier predictions of residual stresses were based on the assumption of elastic unloading. According to Milligan et al [8], the high strength gun steel has a very high Bauschinger effect. To account for the compressive yield strength reduction due to the Bauschinger effect, Underwood and Kendall [9] have tried to estimate the residual stress distribution and its effect on fatigue crack growth rate. Recently, Chen obtained a closed form solution of residual stresses in autofrettaged tubes based on a theoretical material model

taking both Bauschinger and hardening effects into consideration [10]. The new residual stress distribution is very different from that obtained earlier based on the assumption of elastic unloading. The functional stress intensity method developed for the computation of stress intensity factors of a radial crack initiating from the inner radius of a tube may fail due to the presence of a reverse yielding region near the bore. The objective of this paper is to develop a numerical method to overcome such a difficulty and to study the Bauschinger effect on stress intensity factors for a radially cracked, partially autofrettaged gun tube.

II. THE BAUSCHINGER EFFECT AND RESIDUAL STRESSES. The phenomenon that a material lowers its elastic limit in compression (tension) subsequent to a previous stressing in tension (compression) beyond the elastic limit is called the Bauschinger effect. A quantity representing the magnitude of the Bauschinger effect is the Bauschinger effect factor (BEF). This is defined as the ratio of the yield stress upon reverse loading to the initial yield stress (σ_0). The BEF (f) is a function of percent overstrain (ϵ^P). The graph of BEF vs. percent overstrain obtained by Milligan [8] for a modified 4330 steel having a martensitic structure is shown in Figure 1 which was used by Chen in his computations of residual stresses.

Taking the Bauschinger effect factor f into consideration, Chen obtained the closed form solution of residual stresses in autofrettaged tubes [10]. He assumed a material model which exhibits the stress-strain curve shown in Figure 2 during tensile loading and unloading after overstrain. The assumption of elastic-perfectly plastic loading was supported by the fact that very little strain-hardening was observed in the tensile test. A bilinear model for elastic-plastic unloading was assumed since a large slope of strain-hardening ($m'E$) did develop after the occurrence of reverse yielding.

Referring to the point O' in Figure 2 as the origin of a new (primed) coordinate system (σ', ϵ'), the reverse yielding curve can be expressed as

$$\frac{\sigma'}{\sigma_0} = 1 + f + m' \zeta / (1 - m') \quad (1)$$

where $\zeta = (E/\sigma_0)\epsilon'^P$, E is Young's modulus, ϵ'^P is the plastic strain in the primed coordinates. The final residual stress state (denoted by a double prime) is obtained by summing the stress state corresponding to the elastic-perfectly plastic loading and the primed stress state corresponding to unloading. The tangential and radial components can be written as

$$\sigma_\theta'' = \sigma_\theta + \sigma_\theta' \quad , \quad \sigma_r'' = \sigma_r + \sigma_r' \quad (2)$$

Elastic Plastic Loading

Let a and b be the inner and outer radii of the cylinder, respectively, and let the material be elastic-perfectly plastic, obeying the Tresca's yield criterion, the stress components are given by [2]

$$\frac{\sigma_r}{\sigma_0} = \frac{1}{2} \left(\mp 1 + \frac{\rho^2}{b^2} \right) - \ln \frac{\rho}{r} \quad , \quad a \leq r \leq \rho \quad (3)$$

$$\frac{\sigma_r}{\sigma_0} = \frac{1}{2} \left(\frac{\rho^2}{b^2} + \frac{\rho^2}{r^2} \right), \quad \rho \leq r \leq b \quad (4)$$

where ρ is the elastic-plastic boundary relating to the internal pressure p by

$$\frac{p}{\sigma_0} = \frac{1}{2} \left(1 - \frac{\rho^2}{b^2} \right) + \ln \left(\frac{\rho}{a} \right) \quad (5)$$

The equivalent plastic strain can be calculated by

$$\frac{E}{\sigma_0} \epsilon^p = \beta_1 \left(\frac{\rho^2}{r^2} - 1 \right), \quad a \leq r \leq \rho \quad (6)$$

where

$$\beta_1 = \frac{2}{\sqrt{3}} (1 - \nu^2) \quad (7)$$

In Eq. (7) ν is the ratio of Poisson.

Elastic-Plastic Unloading

If the internal pressure p is subsequently removed completely, the unloading may be either elastic or elastic-plastic depending on whether the magnitude of p is less than or greater than p_m which is the minimum pressure to cause the reverse yielding to occur. If $p \leq p_m$, the unloading is entirely elastic and the stress components are

$$\frac{\sigma_r'}{\sigma_0} = \frac{p}{b^2} \left[\pm \frac{b^2}{r^2} - 1 \right] \quad (8)$$

Using Tresca's criterion subject to $\sigma_r'' > \sigma_z'' > \sigma_\theta''$, the reverse yielding will not occur if

$$\sigma_r'' - \sigma_\theta'' \leq f \sigma_0 \quad (9)$$

Substituting Eqs. (2), (3), and (8) into (9), we obtain the expression for p_m

$$\frac{p_m}{\sigma_0} = \frac{1}{2} (1+f) \left(1 - \frac{a^2}{b^2} \right) \quad (10)$$

If $p > p_m$, reverse yielding will occur in a region $a \leq r < \rho'$ with $\rho' < \rho$ upon unloading and we have from Eq. (1)

$$\sigma_r'' - \sigma_\theta'' = f \sigma_0 + m' E \epsilon' P / (1 - m') \quad \text{in} \quad a \leq r < \rho' \quad (11)$$

The stresses in the plastic and elastic regions are given by [3]

$$\sigma_r'/\sigma_0 = p/\sigma_0 - \frac{1}{2} \beta_2'(1+f) \left(\frac{\rho'}{a}\right)^2 \left(1 - \frac{a^2}{r^2}\right) - (1-\beta_2')(1+f) \ln\left(\frac{r}{a}\right), \quad (a \leq r \leq \rho') \quad (12)$$

$$\sigma_\theta'/\sigma_0 = \sigma_r'/\sigma_0 - (1+f) - m' \zeta'/(1-m'), \quad (a \leq r \leq \rho') \quad (13)$$

$$\sigma_r'/\sigma_0 = \frac{1}{2} (1+f) \left[\pm \left(\frac{\rho'}{r}\right)^2 - \left(\frac{\rho'}{b}\right)^2 \right], \quad (\rho' \leq r \leq b) \quad (14)$$

$$\sigma_\theta'/\sigma_0 = \frac{1}{2} (1+f) \left[\pm \left(\frac{\rho'}{r}\right)^2 - \left(\frac{\rho'}{b}\right)^2 \right], \quad (\rho' \leq r \leq b) \quad (15)$$

where

$$\beta_1' = (1-m')/[m' + (1-m')/\beta_1] \quad (16)$$

$$\beta_2' = m' \beta_1'/(1-m') \quad (17)$$

$$\zeta' = \beta_1'(1+f)(\rho'^2/r^2 - 1) \quad (18)$$

The value of ρ' can be found from continuity of σ_r' at ρ' from Eqs. (12) and (14). Figure 3 shows different residual hoop stress distributions for a hollow cylinder with a wall ratio of two and $E/\sigma_0 = 200$, $\nu = 0.3$, $\rho/a = 1.6$ and 2.0. The broken lines are elastic unloading solutions ($f=1$), while solid lines are elastic-plastic unloading with $m' = 0$ and 0.3. It shows a drastic difference in σ_θ'' in the reverse yielding region near the bore and ρ' varies slightly with ρ and m' .

III. FUNCTIONAL STRESS INTENSITY METHOD. We have developed the functional stress intensity method [11] for the computation of stress intensity factors of radial cracks in a pressurized and partially autofrettaged cylinder by combining the finite element method and the weight function method. A weight function vector \underline{h} is a universal function which depends only on geometry and not on loadings [12]. For a given radially cracked ring, if $K_I^{(1)}$, the mode I stress intensity factor, and $v^{(1)}$, the normal component of crack face displacement associated with a symmetric load system 1 are known, then the normal component of the crack face weight functions can be expressed by

$$h_{Iy}(x, \ell) = \frac{H}{2K_I^{(1)}} \frac{\partial v^{(1)}(x, \ell)}{\partial \ell} \quad (19)$$

where $H = E$ for plane stress and $H = E/(1-\nu^2)$ for plane strain, $x = (r-a)$ is a distance measured from the base of the crack along the crack face toward the crack tip, and ℓ is the crack length. For any symmetric load applied to the same cracked ring, the stress intensity factor associated with the new load can be found from

$$K_I = \frac{H}{K_I^{(1)}} \int_0^\ell p(x) \frac{\partial v^{(1)}}{\partial \ell} dx \quad (20)$$

where $p(x)$ is the normal stress in the tangential direction at the crack site due to the new load applied to the uncracked ring. If the uniform tension p_0 at the outer radius is taken as load one, we can find numerical values of $K_I(1)$ and $v(1)$ by the finite element method. We wish to use Eq. (20) to obtain K_I associated with the autofrettage residual stresses. It can be seen from Eqs. (2), (3), (8), and (12) through (15) that the residual stress σ_θ'' has the general expression

$$\sigma_\theta''(r) = A_1 + A_2/r^2 + A_3 \ln(r) \quad (21)$$

for a tube subject to an elastic-plastic loading prior to an elastic unloading, where A_i , $i = 1, 2, 3$ are superposition constants.

Let $p(x)$ in Eq. (20) be σ_θ'' above. The integration to find K_I requires an expression for $\partial v(1)/\partial \ell$. A method which assumes $v(1)$ as a conic section [13] has been used by Grandt [14]. Another method which avoids the need of $\partial v(1)/\partial \ell$ has been developed by Pu. He uses finite elements to obtain functional stress intensities $K_C(1)$, $K_C(r^{-2})$, and $K_C(\ln r)$ defined by

$$K_C(p) = \frac{H}{K(1)} \int_0^\ell p \frac{\partial v(1)}{\partial \ell} dx \quad (22)$$

for a crack face loading $p = 1$, $p = r^{-2}$, and $p = \ln(r)$, respectively. Once the functional stress intensities are known, then the stress intensity factor $K_I(p)$ can be found for a general residual stress distribution of Eq. (21) by an algebraic equation

$$K_I(p) = A_1 K_C(1) + A_2 K_C(r^{-2}) + A_3 K_C(\ln r) \quad (23)$$

This method was successfully applied to various residual stresses predicted from various material models [15].

Now for a radial crack in a reverse yielding zone, the stress intensity factor can still be found by the functional stress intensity method since the residual stress remains the same form as Eq. (21) with the superposition constants obtainable from Eqs. (3) and (13). However, if the radial crack is longer than the reverse yielding zone, the functional stress intensity method fails since the crack face loading $p(x)$ has two different expressions in two different regions: $0 \leq x \leq \rho' - a$ and $\rho' - a \leq x \leq \ell$. Let $p_Y(x)$ and $p_e(x)$ be the crack face loading (residual hoop stress) in the reverse yielding region and the elastic region. Equation (20) becomes

$$K_I = 2 \int_0^{\rho' - a} p_Y(x) h_{IY}(x, \ell) dx + 2 \int_{\rho' - a}^\ell p_e(x) h_{IY}(x, \ell) dx \quad (24)$$

To use this equation, we have to find the explicit crack face weight function h_{IY} using the stiffness derivative finite element technique explained in the next section.

IV. EXPLICIT CRACK FACE WEIGHT FUNCTION. The stiffness derivative method for determining linear elastic stress intensity factors was introduced by Parks [16]. The method was extended to calculate the weight function vector field by Parks et al [17]. An efficient finite element evaluation of explicit weight functions has been established by Sha [18] who combines the stiffness derivative technique with special singular crack tip elements. He applied the method to radial crack problems of a hollow disk [19]. The degenerated quarter-point quadratic elements were used around a crack tip. These singular elements were surrounded by the standard eight-node quadrilateral elements. The virtual crack extension of an amount $\delta\ell$ was simulated by advancing the crack-tip node by $\delta\ell$ in the direction colinear with the mode I radial crack (x-direction). The surrounding quarter-point nodes were also shifted to new locations and there was no change in location of all other nodes. Hence only a few crack-tip elements have experienced changes in elemental stiffness due to the virtual crack extension. This makes the stiffness derivative technique very efficient from a computational viewpoint.

From the displacement form of the finite element method, the equation of equilibrium is

$$[K]\{u\} = \{F\} \quad (25)$$

where $[K]$ is the global stiffness matrix, $\{u\}$ and $\{F\}$ are displacement vector and load vector, respectively. The change in displacement per unit crack extensions can be found by differentiating Eq. (25) with respect to crack length ℓ

$$\frac{d\{u\}}{d\ell} = [K]^{-1} \left[\frac{d\{F\}}{d\ell} - \frac{d[K]}{d\ell} \{u\} \right] \quad (26)$$

Note that $d\{F\}/d\ell = 0$ if we select a load system which consists of only surface tractions not applied on the crack face. The global stiffness derivative $d[K]/d\ell$ is the sum of N_c elemental stiffness derivatives

$$\frac{d[K]}{d\ell} = \sum_{i=1}^{N_c} \frac{d[k_i]}{d\ell} \quad (27)$$

where N_c is the number of crack tip elements and $[k_i]$ is the element stiffness of the i^{th} crack tip element. For a small crack extension $\delta\ell$, the element stiffness derivative may be approximated by

$$\frac{d[k_i]}{d\ell} = \frac{[k_i]_{\ell+\delta\ell} - [k_i]_{\ell}}{\delta\ell} \quad (28)$$

where $[k_i]_{\ell+\delta\ell}$ and $[k_i]_{\ell}$ are element stiffness after and before virtual crack extension, respectively.

The displacement vector is a function of position (x,y) and the crack length ℓ

$$\{u\} = \{u(x, y, \ell)\} \quad (29)$$

Applying the chain rule of differentiation of Eq. (29) with respect to ℓ gives

$$\frac{\partial \{u\}}{\partial \ell} = \frac{d \{u\}}{d \ell} = \frac{\partial \{u\}}{\partial x} \frac{dx}{d \ell} + \frac{\partial \{u\}}{\partial y} \frac{dy}{d \ell} \quad (30)$$

For a mode I crack lying on the x-axis with colinear virtual crack extension, we have $dx/d\ell = 1$ and $dy/d\ell = 0$. Definitions of global coordinates and displacements for isoparametric elements are

$$\begin{aligned} x &= \sum N_i(\xi, \eta) x_i, & u &= \sum N_i(\xi, \eta) u_i \\ y &= \sum N_i(\xi, \eta) y_i, & v &= \sum N_i(\xi, \eta) v_i \end{aligned} \quad (31)$$

where ξ, η are local coordinates; x, y are global coordinates; x_i, y_i are global coordinates of node i ; u_i and v_i are x and y components of displacement at node i ; $N_i(\xi, \eta)$ are shape functions which interpolate the displacement over the element. The finite element evaluation of $\partial \{u\} / \partial x$ can be carried out from

$$\frac{\partial \{u\}}{\partial x} = \frac{1}{\det[J]} \left\{ \frac{\partial [N_i u_i]}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial [N_i u_i]}{\partial \eta} \frac{\partial y}{\partial \xi} \right\} \quad (32)$$

where $[J]$ is the Jacobian matrix. This leads to the expression for the y component of mode I weight function vector at (x, y)

$$h_{Iy}(x, y, \ell) = \frac{H}{K_I(1)(\ell)} \left[\frac{dv(1)}{d\ell} + \frac{1}{\det[J]} \left\{ \frac{\partial [N_i u_i]}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial [N_i u_i]}{\partial \eta} \frac{\partial y}{\partial \xi} \right\} \right] \quad (33)$$

For a hollow disk of $b/a = 2$, Sha and Yang [19] have obtained explicit weight function for a single bore or rim radial crack. In private communication, Sha has provided the following expression to approximate the dimensionless crack face weight function component

$$2\sqrt{\ell} h_{Iy}\left(\frac{r_s}{\ell}\right) = \sum_{n=1}^4 D_n \left(\frac{r_s}{\ell}\right)^{n/2-1} \quad (34)$$

where r_s , the distance from the crack tip along the crack face, is related to x and r by

$$r_s = \ell - x = \ell + a - r \quad (35)$$

The coefficients D_n , $n = 1, 2, 3, 4$, determined by the least square technique, are given in the following table for various crack lengths.

TABLE 1. LEAST SQUARE FITTED COEFFICIENTS OF EQUATION (34) FOR A SINGLE BORE CRACK IN A HOLLOW CYLINDER OF $b/a = 2$ WITH VARIOUS CRACK LENGTHS $c = \ell/(b-a)$

$\frac{\ell}{b-a}$	D_1	D_2	D_3	D_4
0.01	0.7966D+00	-0.7721D-02	0.2935D+00	0.3620D+00
0.02	0.7976D+00	-0.6980D-02	0.2840D+00	0.3608D+00
0.03	0.8130D+00	-0.7176D-01	0.3711D+00	0.3134D+00
0.05	0.8008D+00	-0.5811D-02	0.2649D+00	0.3551D+00
0.06	0.8335D+00	-0.1714D+00	0.4942D+00	0.2421D+00
0.08	0.8051D+00	-0.2789D-01	0.3116D+00	0.3270D+00
0.10	0.8286D+00	-0.1818D+00	0.6119D+00	0.1554D+00
0.20	0.8132D+00	-0.8254D-01	0.5769D+00	0.2414D+00
0.30	0.8026D+00	0.1409D-02	0.5881D+00	0.3756D+00
0.40	0.7981D+00	0.4661D-01	0.6908D+00	0.4935D+00
0.50	0.7956D+00	0.7236D-01	0.9065D+00	0.5437D+00
0.60	0.7990D+00	0.4395D-01	0.1361D+01	0.4059D+00
0.70	0.8004D+00	0.2255D-01	0.2130D+01	-0.3915D-01
0.80	0.7976D+00	0.4388D-01	0.3537D+01	-0.1131D+01

V. STRESS INTENSITY FACTORS. To use Eqs. (34) and (24), we need a certain transformation of variables. Let us use a , the bore radius, to normalize all linear lengths and use the same notations (before normalization) to denote the normalized lengths except that a is unit and b is w , the wall ratio. Denote r_s/ℓ by τ , Eq. (24) associated with the residual hoop stress becomes

$$\begin{aligned} \frac{K_I}{\sigma_o \sqrt{\ell}} &= \int_0^1 \left(\frac{\sigma_\theta''}{\sigma_o} \right)_Y \left(\sum_{n=1}^4 D_n \tau^{n/2-1} \right) d\tau + \int_0^{\tau'} \left(\frac{\sigma_\theta''}{\sigma_o} \right)_e \left(\sum_{n=1}^4 D_n \tau^{n/2-1} \right) d\tau \\ &= \int_0^1 \left(\frac{\sigma_\theta''}{\sigma_o} \right)_e \left(\sum_{n=1}^4 D_n \tau^{n/2-1} \right) d\tau + \int_{\tau''}^1 \left[\left(\frac{\sigma_\theta''}{\sigma_o} \right)_Y - \left(\frac{\sigma_\theta''}{\sigma_o} \right)_e \right] \left[\sum_{n=1}^4 D_n \tau^{n/2-1} \right] d\tau \quad (36) \end{aligned}$$

where $(\sigma_\theta''/\sigma_o)_e$ is the residual hoop stress in the elastic unloading region, while $(\sigma_\theta''/\sigma_o)_Y$ is that in the reverse yielding region and τ' is the value of τ corresponding to $r = \rho'$, the elastic-plastic interface during unloading. The residual hoop stress can be represented by the general form

$$\left(\frac{\sigma_\theta''}{\sigma_o} \right)_e = A_{1e} + A_{2e} r^{-2}(\tau) + A_{3e} \ln(r/\tau) \quad (37)$$

$$\left(\frac{\sigma_\theta''}{\sigma_o} \right)_Y = A_{1Y} + A_{2Y} r^{-2}(\tau) + A_{3Y} \ln(r/\tau) \quad (38)$$

where the coefficients A_{1e}, \dots, A_{3Y} , can be obtained from Eqs. (2), (3), and (12) through (15). From Eq. (35) and the definition of τ , we have

$$r(\tau) = \ell(\ell_0 - \tau) \quad , \quad \ell_0 = (1+\ell)/\ell \quad (39)$$

Denoting $I_{1n}(\alpha, \beta)$, $I_{2n}(\alpha, \beta)$, and $I_{3n}(\alpha, \beta)$ as the following definite integrals,

$$\begin{aligned} I_{1n}(\alpha, \beta) &= \int_{\alpha}^{\beta} \tau^{n/2-1} d\tau \\ I_{2n}(\alpha, \beta) &= \int_{\alpha}^{\beta} r^{-2}(\tau) \tau^{n/2-1} d\tau \quad n = 1, 2, 3, 4 \\ I_{3n}(\alpha, \beta) &= \int_{\alpha}^{\beta} \ell n(r(\tau)) \tau^{n/2-1} d\tau \end{aligned} \quad (40)$$

Eq. (36) can be written in the form

$$\frac{K_I}{\sigma_0 \sqrt{\pi \ell}} = \frac{1}{\sqrt{\pi}} \left[\sum_{i=1}^3 A_{ie} \sum_{n=1}^4 D_n I_{in}(0, 1) + \sum_{i=1}^3 (A_{iY} - A_{ie}) \sum_{n=1}^4 D_n I_{in}(\tau', 1) \right] \quad (41)$$

The expressions of integrals $I_{in}(\alpha, \beta)$, $i = 1, 2, 3$, $n = 1, \dots, 4$, are easy to carry out and are omitted. Equation (41) can be used for any loading which yields a hoop stress of the form given by Eq. (21) and for any multiply-cracked cylinder as long as the coefficients D_n are known for that particular cracked geometry. In case there is no reverse yielding region, Eq. (41) reduces to

$$\frac{K_I}{\sigma_0 \sqrt{\pi \ell}} = \frac{1}{\sqrt{\pi}} \sum_{i=1}^3 A_{ie} \sum_{n=1}^4 D_n I_{in}(0, 1) \quad (42)$$

For a shallow crack which lies entirely in the reverse yielding region, Eq. (41) becomes

$$\frac{K_I}{\sigma_0 \sqrt{\pi \ell}} = \frac{1}{\sqrt{\pi}} \sum_{i=1}^3 A_{iY} \sum_{n=1}^4 D_n I_{in}(0, 1) \quad (43)$$

Table 1 gives D_n for single bore crack of various discrete crack lengths. For a crack length not given in Table 1, Sha [19] suggested the use of cubic spline interpolation to obtain the weight function for that length from discrete values in Table 1 followed by the least square technique to calculate D 's.

Equation (42) is used to compute $K_I/p_0 \sqrt{\pi \ell}$ for a single bore crack in a hollow cylinder of $w = 2$ subject to uniform tension p_0 at the outside cylindrical surface. The Lamé solution gives $A_1 = A_2 = w^2/(w^2-1)$ and $A_3 = 0$.

Numerical results for various crack lengths are very accurate in comparison with results previously reported in [20]. Further check of numerical computations of $K_I/\sigma_0\sqrt{\pi l}$ from Eq. (42) is done for a partially autofrettaged cylinder without reverse yielding. The results confirm those published in [21].

For a cylinder of $w = 2$, the Bauschinger effect factors for 60 percent, 80 percent, and 100 percent overstrain are, from Figure 1, $f = 0.44$, 0.42 , and 0.38 , respectively. Using $\nu = 0.3$, $E/\sigma_0 = 200$, the elastic-plastic interface ρ' during unloading depends on f and m' . It varies from $\rho' = 1.106$ for $\epsilon = 0.6$ ($f = 0.44$) and $m' = 0.3$ to $\rho' = 1.20$ for $\epsilon = 1$ ($f = 0.38$) and $m' = 0$. The residual hoop stress distributions near the bore for these two cases are shown in solid lines in Figure 3 while broken lines are corresponding stresses without taking Bauschinger effect into consideration. The residual stress in the reverse yielding region varies drastically with f and m' . Stress intensity factors due to residual stresses are calculated from Eq. (43) or (41) depending on whether the crack tip is inside or outside of the reverse yielding zone. Figures 4 and 5 are graphs of $K_I/(\sigma_0\sqrt{\pi l})$ as a function of dimensionless crack length $c = l/(w-1)$ for $\epsilon = 0.6$ and $\epsilon = 1$, respectively. Superposing stress intensity factors due to an internal pressure $p = \sigma_0/L_f$, the combined stress intensity factors from both p and residual stresses are shown in Figures 6 and 7 for $L_f = 3$ and for $\epsilon = 0.6$ and $\epsilon = 1.0$, respectively. Figure 8 is a similar graph for $L_f = 1.5$ and $\epsilon = 1.0$. In Figures 6, 7, and 8, the curves in the centerline correspond to $m' = 0$, the curves in the broken line are for $m' = 0.3$, while the solid lines are results obtained earlier without taking reverse yielding into consideration. For $c = 0.01$, Figure 7, values of $K_I/(\sigma_0\sqrt{\pi l})$ are 0.55 and 0.37 for $m' = 0$ and 0.3 , respectively, versus the corresponding value 0.058 from the solid line. It indicates that the Bauschinger effect will greatly reduce the advantageous effect of compressive residual hoop stress introduced by the autofrettage process.

VI. CONCLUSION. The Bauschinger effect of the high strength gun steel causes the presence of the reverse yielding in a partially autofrettaged cannon tube of wall ratio of two. This reverse yielding reduces the magnitude of compressive hoop stress considerably in the reverse yielding region. This stress reduction will result in much higher stress intensity factors in a shallow bore crack in a pressurized and autofrettaged gun tube. The higher the stress intensity factor implies the lower the fatigue life.

The expression in Eq. (34) for explicit crack face weight function obtained by a combination of stiffness derivative technique and special singular crack tip elements is highly accurate. It recovers the previous stress intensity factor results for a single bore crack in a tube subject to various loading conditions. The method can treat any crack face loading including stress discontinuities and stress gradient discontinuities over the crack face.

Numerical results are limited to single bore radial cracks in this paper. However, the method is general for either bore cracks or rim cracks for any number of cracks as long as we have obtained the explicit crack face weight function for that particular cracked geometry.

REFERENCES

1. Davidson, T. E., Throop, J. F., and Underwood, J. H., "Failure of a 175 mm Cannon Tube and the Resolution of the Problem Using an Autofrettage Design," Case Studies in Fracture Mechanics, T. P. Rich and D. J. Cartwright, Eds., AMMRC MS77-5, Army Materials and Mechanics Research Center, 1977, pp. 1-13.
2. Koiter, W. T., "On Partially Plastic Tubes," C. B. Biezeno Anniversary Volume on Applied Mechanics, N.V. de Technische Uitgeverij, H. Stam. Haarlem, 1953.
3. Bland, D. R., "Elastoplastic Thick-Walled Tubes of Work-Hardening Materials Subject to Internal and External Pressures and Temperature Gradients," *Journal of Mechanics and Physics of Solids*, Vol. 4, 1956, pp. 209-229.
4. Chen, P. C. T., "The Finite Element Analysis of Elastic-Plastic Thick-Walled Tubes," *Proceedings of Army Symposium on Solid Mechanics, The Role of Mechanics in Design-Ballistic Problems*, 1972, pp. 243-253.
5. Chen, P. C. T., "Numerical Prediction of Residual Stresses in an Autofrettaged Tube of Compressible Material," *Proceedings of the 1981 Army Numerical Analysis and Computer Conference*, pp. 351-362.
6. Capsimalis, G. P., Haggerty, R. F., and Loomis, K., "Computer Controlled X-Ray Stress Analysis for Inspection of Manufactured Components," Technical Report WVT-TR-77001, Watervliet Arsenal, Watervliet, NY, 1977.
7. Frankel, J., Scholz, W., Capsimalis, G., and Korman, W., "Residual Stress Measurement in Circular Steel Cylinder," ARDC Technical Report ARLCB-TR-84018, Benet Weapons Laboratory, Watervliet, NY, 1984.
8. Milligan, R. V., Koo, W. H., and Davidson, T. E., "The Bauschinger Effect in a High-Strength Steel," *Journal of Basic Engineering, Transaction ASME*, Vol. 88, 1966, pp. 480-488.
9. Underwood, J. H. and Kendall, D. P., "Fracture Analysis of Thick-Wall Cylinder Pressure Vessels," *Theoretical and Applied Fracture Mechanics*, Vol. 2, 1984, pp. 47-58.
10. Chen, P. C. T., "The Bauschinger and Hardening Effect on Residual Stresses in an Autofrettaged Thick-Walled Cylinder," to appear in *Journal of Pressure Vessel Technology*.
11. Pu, S. L., "A Functional Stress Intensity Approach to Multiply Cracked, Partially Autofrettaged Cylinders," *Transactions of the Twenty-Eighth Conference of Army Mathematicians*, ARO Report 83-1, 1983, pp. 263-283.
12. Rice, J. R., "Some Remarks on Elastic Crack-Tip Stress Fields," *Int. Journal of Solids and Structures*, Vol. 8, 1972, pp. 751-758.

13. Orange, T. W., "Crack Shapes and Stress Intensity Factors For Edge-Cracked Specimens," ASTM STP 513, 1972, pp. 71-78.
14. Grandt, A. F., "Two Dimensional Stress Intensity Factor Solutions For Radially Cracked Rings," Technical Report AFML-TR-75-121, Air Force Materials Laboratory, 1975.
15. Pu, S. L. and Chen, P. C. T., "Stress Intensity Factors For Radial Cracks in a Prestressed Thick-Walled Cylinder of Strain-Hardening Materials," Journal of Pressure Vessel Technology, Vol. 105, No. 2, 1983, pp. 117-123.
16. Parks, D. M., "A Stiffness Derivative Finite Element Technique For Determination of Crack Tip Stress Intensity Factors," International Journal of Fracture, Vol. 10, No. 4, 1974, pp. 487-502.
17. Parks, D. M. and Kamenetzky, E. M., "Weight Functions From Virtual Crack Extension," International Journal for Numerical Methods in Engineering, Vol. 14, 1979, pp. 1693-1706.
18. Sha, George T., "Stiffness Derivative Finite Element Technique to Determine Nodal Weight Functions With Singularity Elements," Engineering Fracture Mechanics, Vol. 19, No. 4, 1984, pp. 685-699.
19. Sha, George T. and Yang, Chien-Tung, "Weight Functions of Radial Cracks in Hollow Disks," Presented at ASME 1985 Gas Turbine Conference in Houston, Texas, 1985.
20. Pu, S. L. and Hussain, M. A. "Stress Intensity Factors For a Circular Ring With Uniform Array of Radial Cracks Using Cubic Isoparametric Singular Elements," Fracture Mechanics, ASTM STP 677, 1979, pp. 685-699.
21. Pu, S. L. and Hussain, M. A., "Stress Intensity Factors For Radial Cracks in a Partially Autofrettaged Thick-Wall Cylinder," Fracture Mechanics, Fourteenth Symposium - Vol. 1: Theory and Analysis, ASTM STP 791, 1983, pp. I-194-I-215.

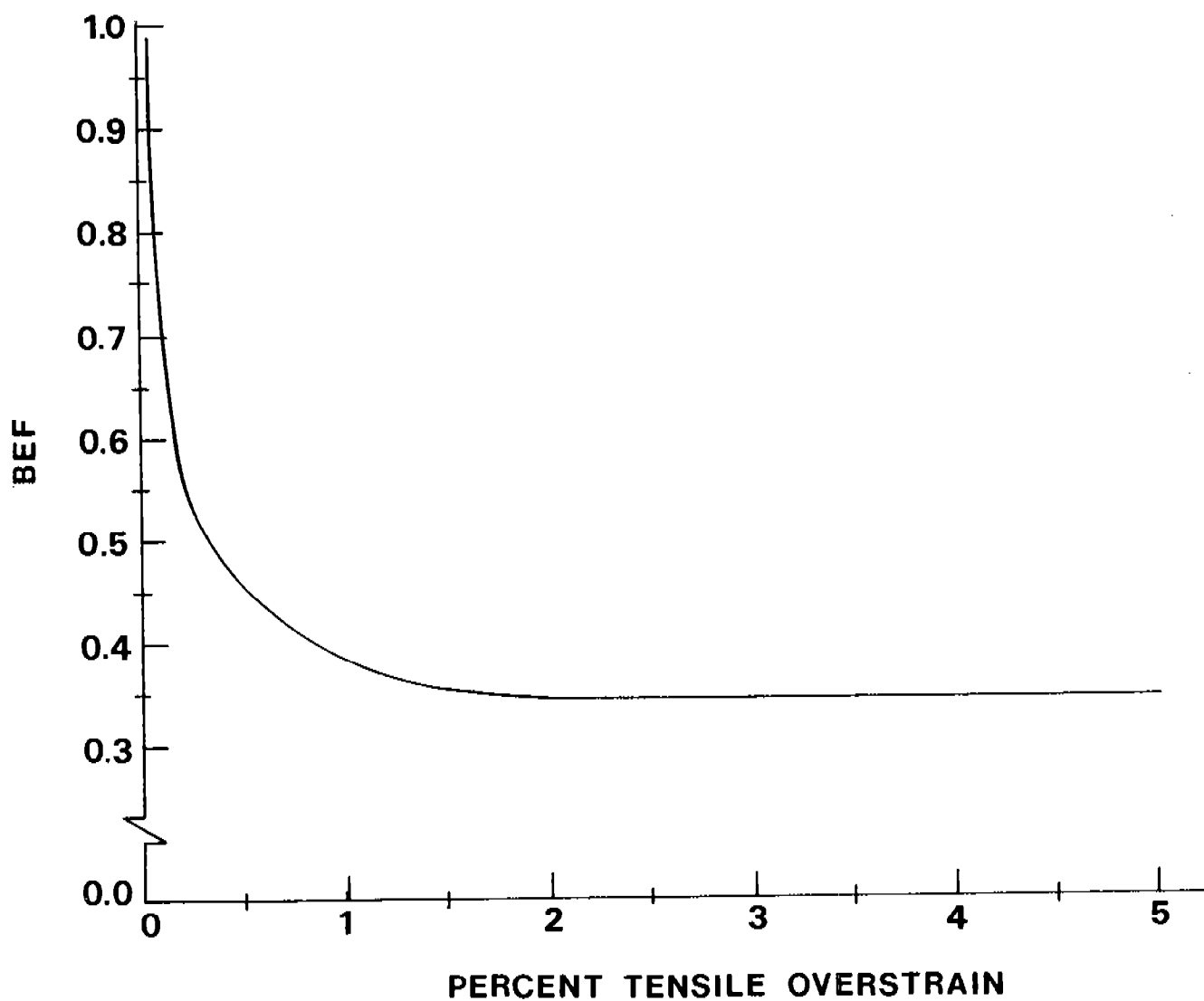


FIGURE 1

Bauschinger effect factor vs. percent tensile overstrain, martensitic structure
of a 4330 modified steel

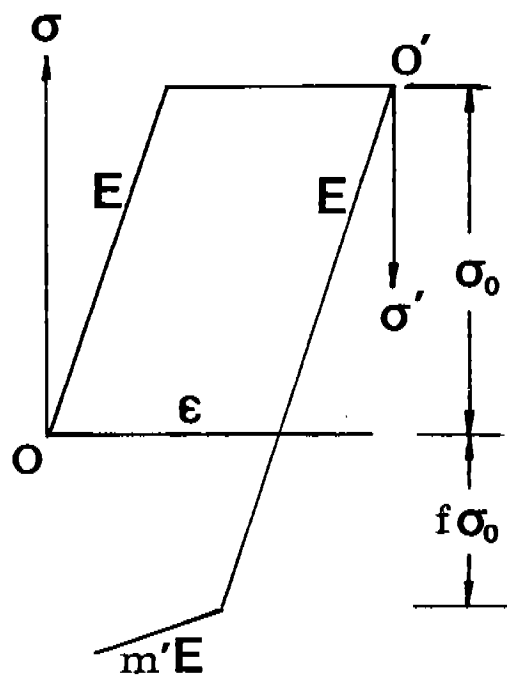


FIGURE 2

Stress-strain curves during loading and unloading

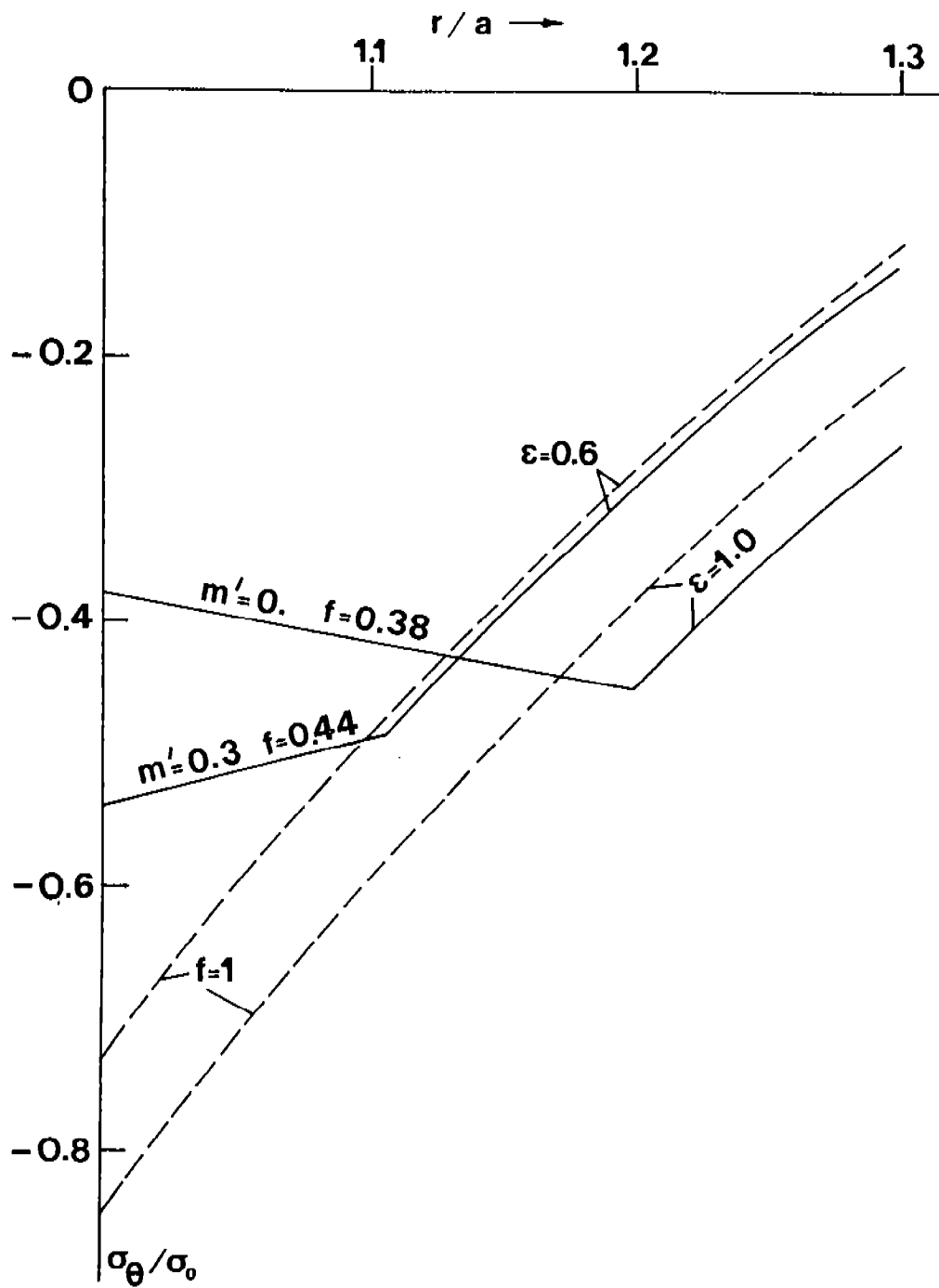


FIGURE 3

Dependence of σ_θ/σ_0 on f and m' in the reverse yielding region of a cylinder of wall ratio of two

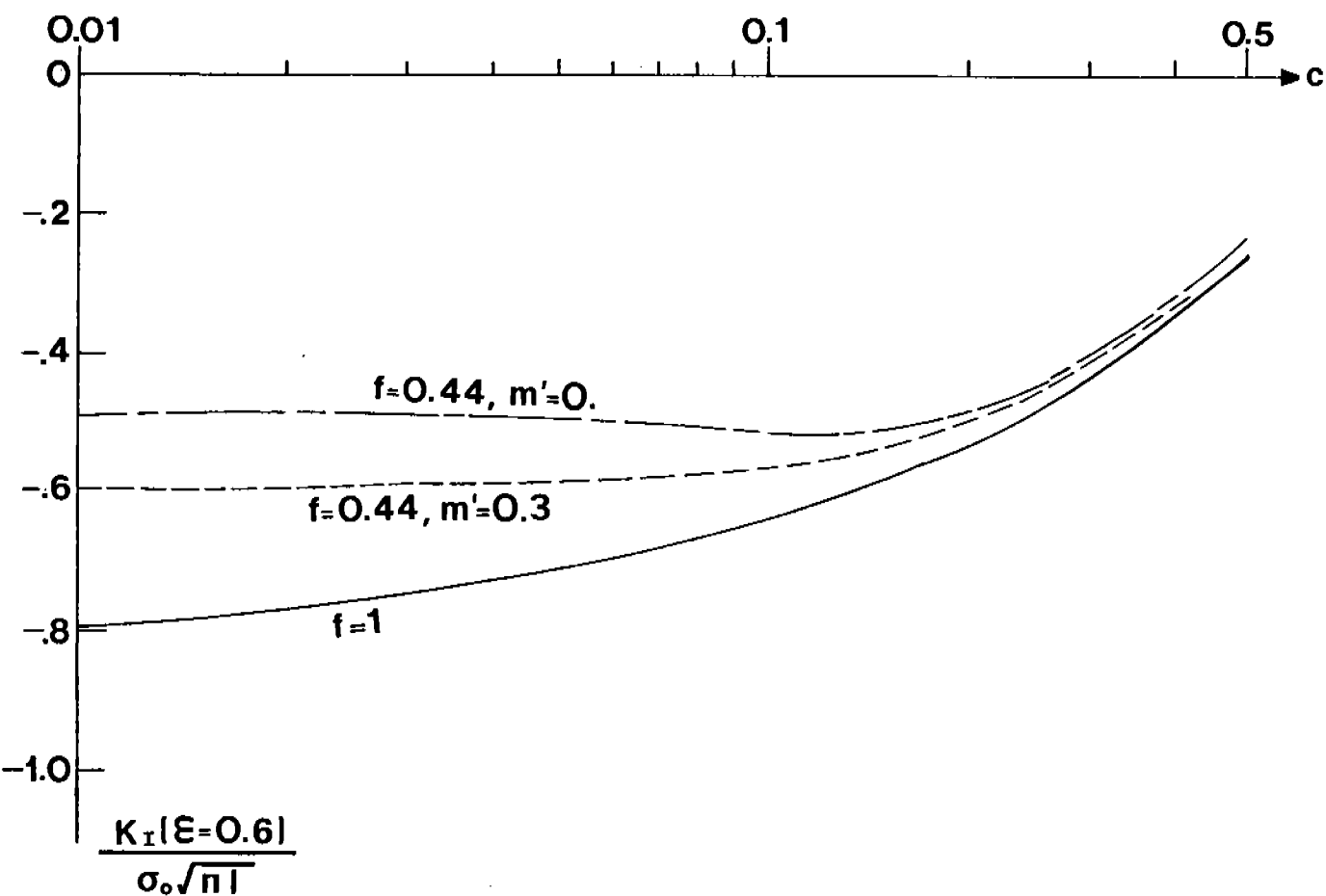


FIGURE 4

Dimensionless stress intensity factors as a function of dimensionless crack length c due to compressive hoop stress in a cylinder having 60 percent degree of autofretage for various values of f and m'

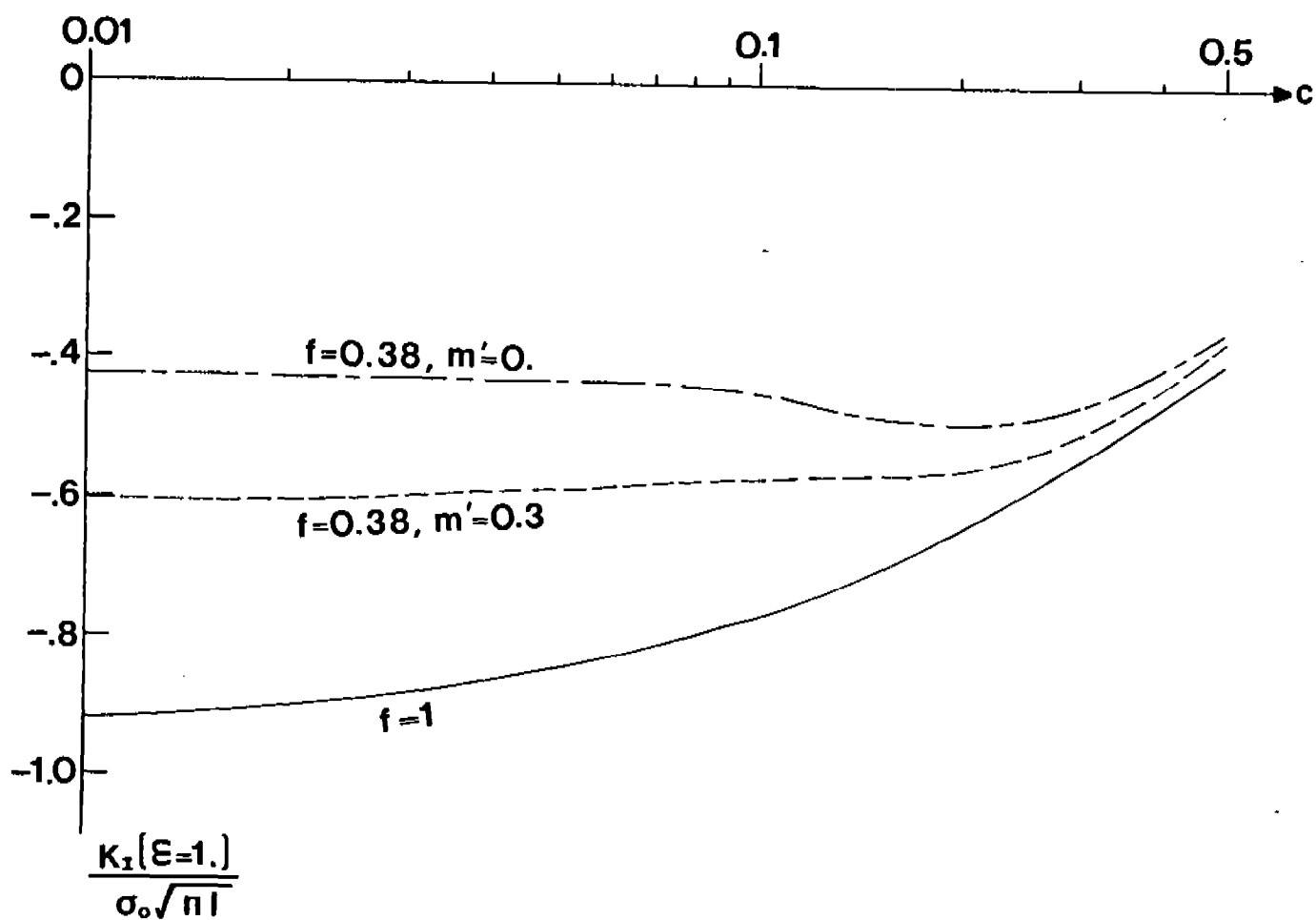


FIGURE 5

Dimensionless stress intensity factors as a function of dimensionless crack length c due to compressive hoop stress in a cylinder having 100 percent degree of autofretage for various values of f and m'

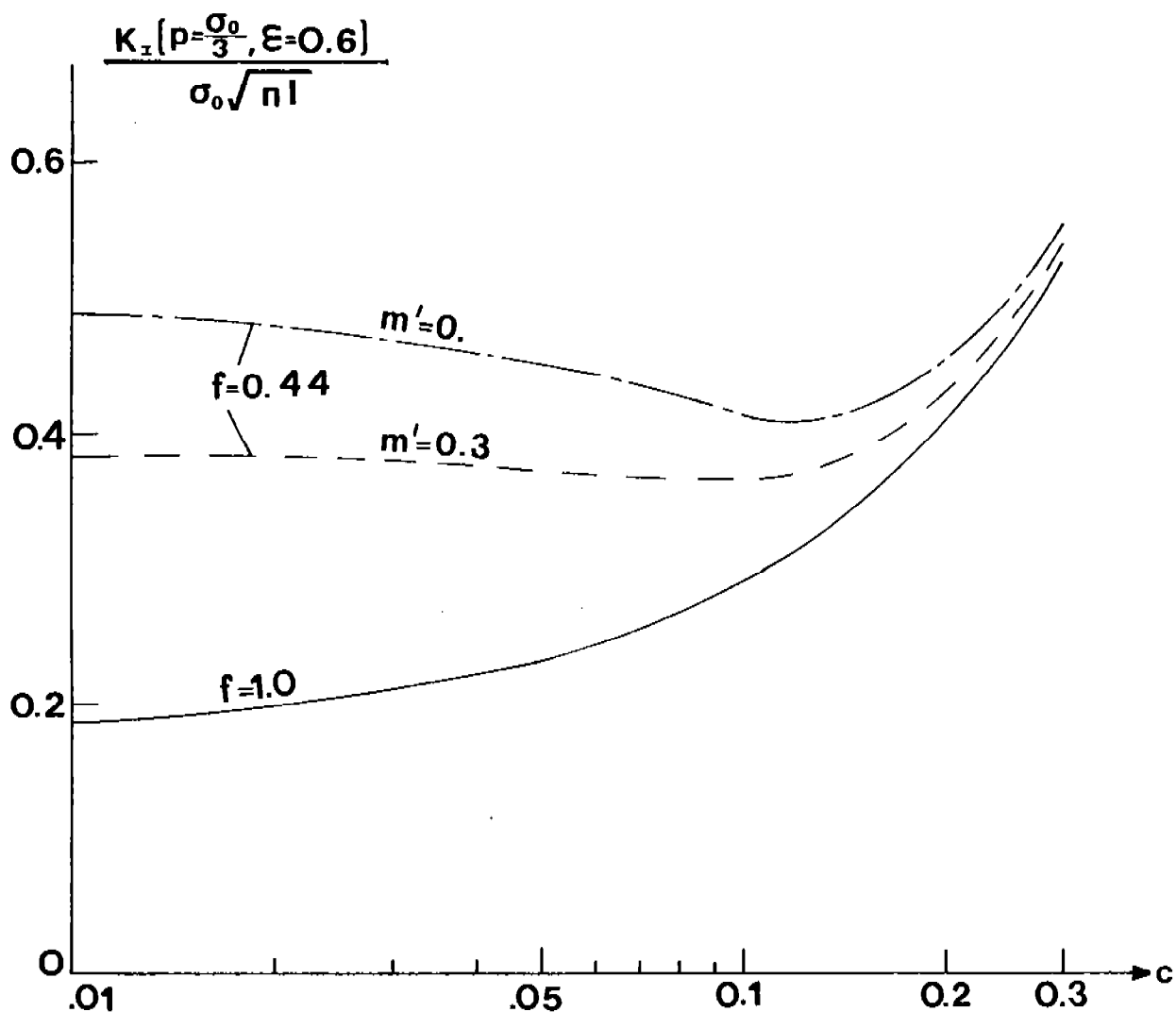


FIGURE 6

Resultant stress intensity factors as a function of crack length c for various values of f and m' in a pressurized, autofrettaged cylinder. Internal pressure

$$p = \sigma_0/3 \text{ and degree of autofrettage } \epsilon = 0.6$$

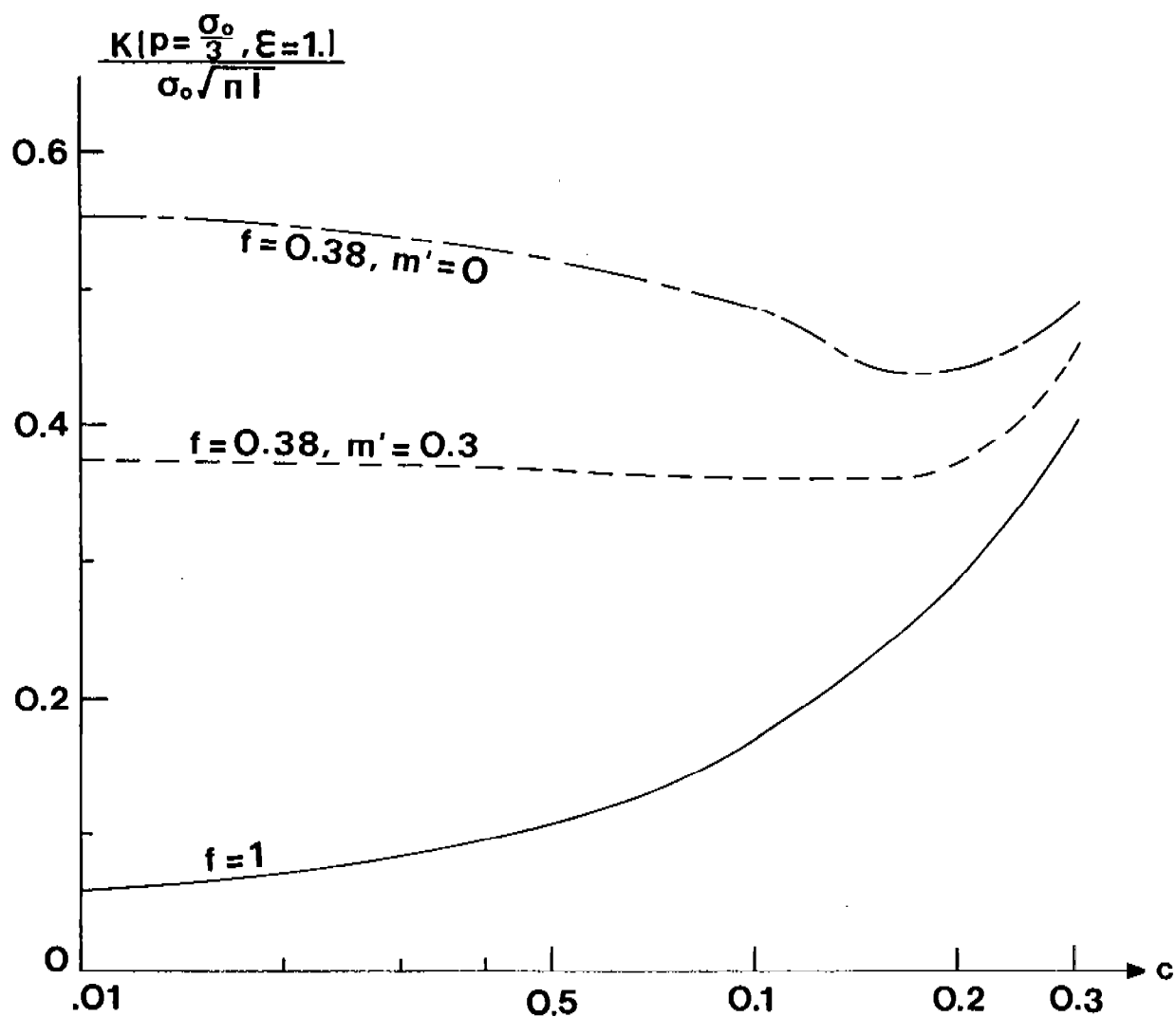


FIGURE 7

Resultant stress intensity factors as a function of crack length c for various values of f and m' in a pressurized, autofrettaged cylinder. Internal pressure

$$p = \sigma_o/3 \text{ and degree of autofrettage } \epsilon = 1.0$$

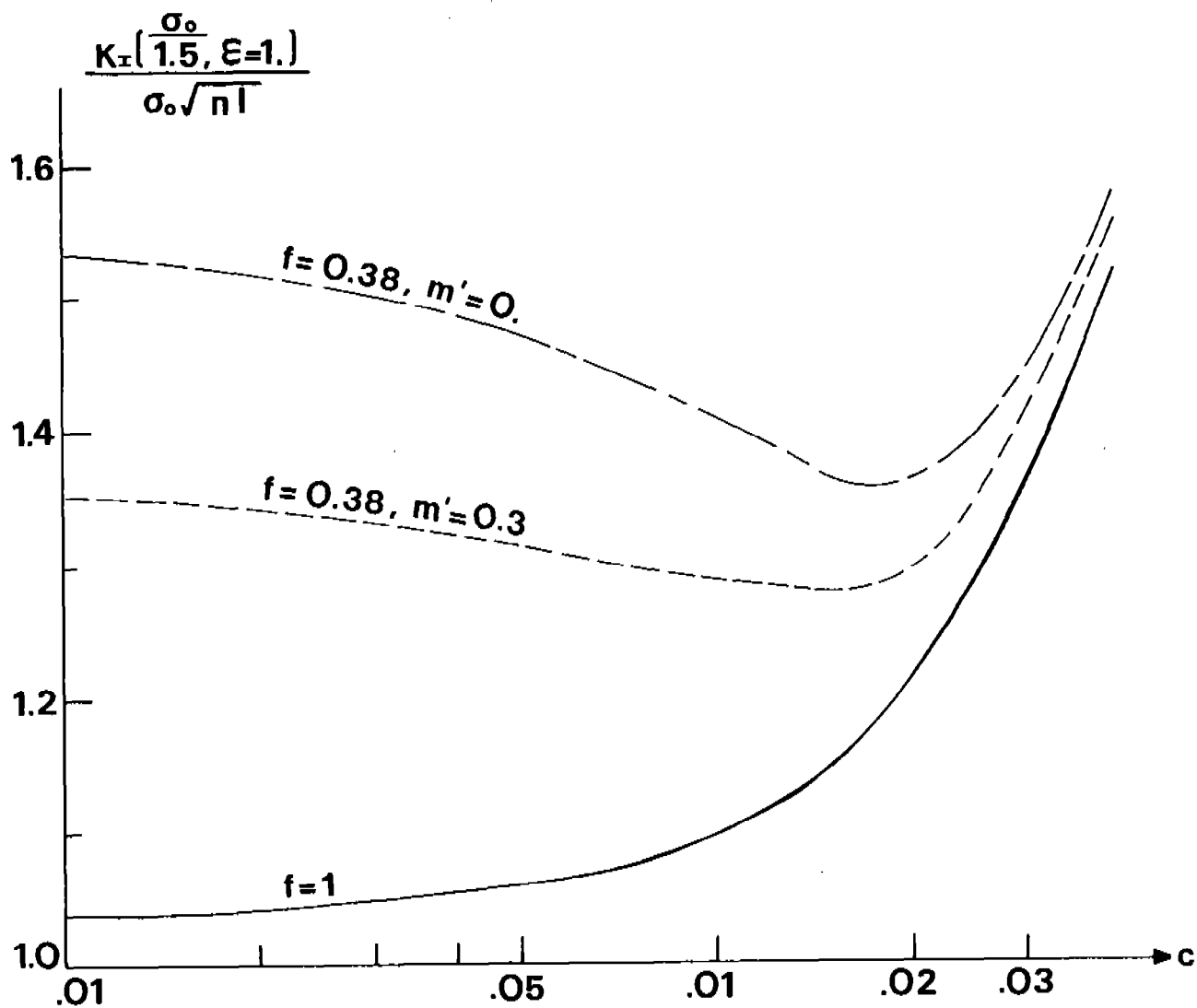


FIGURE 8

Resultant stress intensity factors as a function of crack length c for various values of f and m' in a pressurized, autofrettaged cylinder. Internal pressure

$$p = \sigma_0/1.5 \text{ and degree of autofrettage } \epsilon = 1.0$$

ELASTIC-PLASTIC LOADING AND UNLOADING IN A THICK TUBE WITH KINEMATIC HARDENING THEORY

Peter C. T. Chen

U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189-5000

ABSTRACT. Using Tresca's yield criterion, its associated flow rule, and the linear hardening law, analytical solutions are obtained for elastic-plastic loading and unloading problems in a thick tube subjected to uniform internal pressure. Explicit expressions for the displacement, strains, and stresses are presented and numerical results in a closed-end tube are calculated.

I. INTRODUCTION. The importance of the determination of residual stresses in a prestressed thick-walled cylinder is well-known and elastic-plastic loading response has been extensively studied [1-5]. Most of the earlier solutions for residual stresses were based on the assumption of elastic unloading and only a few considered elastic-plastic unloading [2,5]. Bland's work [2], which neglects the Bauschinger effect, is based on the Tresca's yield condition and isotropic hardening rule. Kinematic hardening is the simplest theory that can model the Bauschinger effect [6,7]. If unloading does not occur, there is no difference between the kinematic and isotropic hardening models. For unloading with reverse yielding, the results based on these two models will be different as shown in a recent paper [5] using the ADINA finite element code [8]. The von Mises' yield condition and its associated flow rules were used in both models.

In this paper a closed-form solution for elastic-plastic loading and unloading in pressurized thick-walled cylinders is presented using Tresca's yield criterion, its associated flow rule, and the linear kinematic hardening law. Numerical results are calculated for a closed-end tube.

II. ELASTIC-PLASTIC LOADING. Consider a thick-walled cylinder, internal radius a and external radius b , which is subjected to internal pressure p . The material is assumed to be elastic-plastic, obeying the Tresca's yield criterion, the associated flow theory, and a linear strain-hardening rule. Using the isotropic hardening theory, the elastic-plastic solution has been obtained by Bland [2]. In order to consider the Bauschinger effect, the kinematic hardening theory is used here. Subject to the condition $\sigma_\theta > \sigma_z > \sigma_r$, Tresca's yield criterion for the Prager-hardening rule [6] states that yielding occurs when

$$(\sigma_\theta - \alpha_\theta) - (\sigma_r - \alpha_r) = K_0 \quad (1)$$

where

$$\alpha_\theta = c \epsilon_\theta^p, \quad \alpha_r = c \epsilon_r^p \quad (2)$$

define the position of the center of the yield surface, c is a material constant and K_0 the initial yield stress. The associated flow rule states that

$$d\epsilon_\theta^p = -d\epsilon_r^p \quad \text{and} \quad d\epsilon_z^p = 0 \quad (3)$$

For the case of linear strain-hardening, the yield stress curve can be represented by a straight line,

$$K/K_0 = 1 + \eta \epsilon^p \quad \text{and} \quad \eta = (E/K_0)^m / (1-m) \quad (4)$$

where η is a material constant and the equivalent plastic strain ϵ^p is defined by

$$\epsilon^p = \sqrt{2/3} \int \{ (d\epsilon_\theta^p)^2 + (d\epsilon_r^p)^2 \}^{1/2} = \frac{2}{\sqrt{3}} \epsilon_\theta^p \quad (5)$$

The elastic-plastic solution for the stresses and displacements can be obtained explicitly. The expressions in the plastic range ($a < r < \rho$) are

$$\sigma_r/K_0 = \frac{1}{2} \left(1 + \frac{\rho^2}{b^2} \right) + \frac{1}{2} \eta \beta \left(\frac{\rho^2}{r^2} - 1 \right) - (1 - \eta \beta) \log \frac{\rho}{r} \quad (6)$$

$$\sigma_\theta/K_0 = \frac{1}{2} \left(1 + \frac{\rho^2}{b^2} \right) + \frac{1}{2} \eta \beta \left(\frac{\rho^2}{r^2} - 1 \right) - (1 - \eta \beta) \log \frac{\rho}{r} \quad (7)$$

$$\sigma_z/K_0 = \nu \rho^2/b^2 - 2\nu(1-\eta\beta) \log \frac{\rho}{r} + E\epsilon_z/K_0 \quad (8)$$

$$(E/K_0)(u/r) = (1-2\nu)(1+\nu)(\sigma_r/K_0) - \nu E\epsilon_z/K_0 + (1-\nu^2)\rho^2/r^2 \quad (9)$$

and in the elastic range ($\rho < r < b$)

$$\sigma_r/K_0 = \frac{1}{2} \left(\frac{\rho^2}{b^2} + \frac{\rho^2}{r^2} \right) \quad (10)$$

$$\sigma_\theta/K_0 = \frac{1}{2} \left(\frac{\rho^2}{b^2} + \frac{\rho^2}{r^2} \right) \quad (11)$$

$$\sigma_z/K_0 = E\epsilon_z/K_0 + \nu \rho^2/b^2 \quad (12)$$

$$(E/K_0)u/r = \frac{1}{2} (1+\nu) [\rho^2/r^2 + (1-2\nu)\rho^2/b^2] - \nu E\epsilon_z/K_0 \quad (13)$$

where

$$E\epsilon_z/K_0 = \frac{(\mu-2\nu)}{(b^2/a^2-1)} (p/K_0) \quad (14)$$

$$\mu = 0 \text{ (open-end)} , \quad 1 \text{ (closed-end)}$$

and

$$\beta^{-1} = \eta + \frac{\sqrt{3}}{2} (E/K_0)/(1-\nu^2) \quad (15)$$

The yield surface moves in translation during plastic deformation as given by

$$\alpha_\theta = -\alpha_r = (\sqrt{3}/2)c\epsilon^P \quad \text{and} \quad \epsilon^P = \beta(\rho^2/r^2-1) \quad (16)$$

The elastic plastic surface ρ is related to the internal pressure p by

$$p/K_0 = \frac{1}{2} (1-\rho^2/b^2) + (1-n\beta)\log \rho/a + \frac{1}{2} n\beta (\rho^2/a^2-1) \quad (17)$$

III. REVERSE YIELDING. If the pressure p given by Eq. (17) is subsequently removed completely with no reverse yielding, the unloading is entirely elastic and the solution is given by

$$\sigma_r' = \frac{p}{b^2/a^2 - 1} \left[\pm \frac{b^2}{r^2} - 1 \right] \quad (18)$$

$$\sigma_\theta' = \frac{p}{b^2/a^2 - 1} \left[\pm \frac{b^2}{r^2} - 1 \right] \quad (19)$$

$$\sigma_z' = \nu(\sigma_r' + \sigma_\theta') + E\epsilon_z' \quad (20)$$

$$E\epsilon_z' = -(\mu-2\nu)p/(b^2/a^2-1) \quad (21)$$

$$Eu'/r = -[(1-\nu - \mu\nu) + (1+\nu)b^2/r^2]p/(b^2/a^2-1) \quad (22)$$

The residual stress system, which will be denoted by two primes, is the sum of the system produced by loading and that produced by unloading, i.e., $\sigma_r'' = \sigma_r + \sigma_r'$, etc. Assuming the kinematic hardening rule and using Tresca's criterion subject to $\sigma_r'' > \sigma_z'' > \sigma_\theta''$, the reverse yielding will not occur if

$$(\sigma_r'' - \alpha_r) - (\sigma_\theta'' - \alpha_\theta) \leq K_0 \quad (23)$$

Substituting the loading and unloading solution into Eq. (23), we can determine the minimum pressure (p_m) for reverse yielding to occur. The equation for p_m is given by

$$p_m/K_0 = (1-a^2/b^2) \quad (24)$$

Equating (17) to (24), we can determine the maximum amount of overstrain for reverse yielding not to occur.

IV. ELASTIC-PLASTIC UNLOADING. Now suppose that the loading has been such that the internal pressure is larger than p_m given by Eq. (24). On unloading, yielding will occur for $a \leq r \leq \rho'$ with $\rho' < \rho$. Using the kinematic hardening rule during unloading and assuming $\sigma_r'' > \sigma_z'' > \sigma_\theta''$, we have

$$(\sigma_r'' - \alpha_r'') - (\sigma_\theta'' - \alpha_\theta'') = K_0 \quad (25)$$

where

$$\alpha_r'' = c\epsilon_r''^P, \quad \alpha_\theta'' = c\epsilon_\theta''^P \quad (26)$$

Since the residual stress system is the sum of two systems produced by loading and unloading, combining Eqs. (1) and (25) leads to

$$(\sigma_r' - \alpha_r') - (\sigma_\theta' - \alpha_\theta') = 2K_0 \quad (27)$$

where

$$\alpha_r' = c\epsilon_r'^P, \quad \alpha_\theta' = c\epsilon_\theta'^P \quad (28)$$

During elastic-plastic unloading, the associated flow theory states that

$$d\epsilon_\theta'^P = -d\epsilon_r'^P < 0 \quad \text{and} \quad d\epsilon_z'^P = 0 \quad (29)$$

It has been assumed that the sign of $d\epsilon_\theta'^P$ is the same throughout the unloading process and that is negative. This will be the case when the internal pressure is removed during unloading. Since $\epsilon_z' = \epsilon_z'^e = -\epsilon_z'$ is known, we can use Hooke's law and the equilibrium equation,

$$d\sigma_r'/dr = (\sigma_\theta' - \sigma_r')/r \quad (30)$$

to express σ_z' in terms of σ_r'

$$\sigma_z' = E\epsilon_z' + 2\nu\sigma_r' + \nu r(d\sigma_r'/dr) \quad (31)$$

Since the dilation is purely elastic

$$(du'/dr) + u'/r + \epsilon_z' = E^{-1}(1-2\nu)(\sigma_r' + \sigma_\theta' + \sigma_z') \quad (32)$$

On integration using Eqs. (30) and (31), we obtain

$$ru' = (1-2\nu)(1+\nu)E^{-1}r^2 \sigma_r' - \nu\epsilon_z'r^2 + A \quad (33)$$

where A is a constant. The strain components can be expressed in terms of σ_r' and $(d\sigma_r'/dr)$. The plastic strain components are

$$\epsilon_\theta'^P = -\epsilon_r'^P = Ar^{-2} - (1-\nu^2)E^{-1}r(d\sigma_r'/dr) \quad (34)$$

Using Eq. (31) together with Eqs. (27) and (28), we have

$$r(d\sigma_r'/dr) = -2(K_0 - c\epsilon_\theta'^P) \quad (35)$$

Substituting Eq. (35) into Eq. (34) and determining the constant A by the condition $\epsilon_\theta'^P = 0$ at ρ' , we obtain

$$A = -2(1-\nu^2)(K_0/E)\rho'^2 \quad (36)$$

and

$$(E/K_0)\epsilon_\theta'^P = -2(\rho'^2/r^2 - 1)/[2c/E + (1-\nu^2)^{-1}] \quad (37)$$

Equations (35) and (36) with the boundary condition at $r = a$ suffice to determine σ_r' in the plastic region. The expressions for the stresses in ($a < r < \rho'$) are given explicitly by

$$\sigma_r'/K_0 = p/K_0 - \eta\beta(\rho'^2/a^2 - \rho'^2/r^2) - 2(1-\eta\beta)\log(r/a) \quad (38)$$

$$\sigma_{\theta}'/K_0 = \sigma_r'/K_0 - 2 - 2\eta\beta(\rho'^2/r^2 - 1) \quad (39)$$

$$\sigma_z'/K_0 = \nu(\sigma_r' + \sigma_{\theta}')/K_0 - E\varepsilon_z/K_0 \quad (40)$$

and in $(\rho' < r < b)$ given by

$$\sigma_r'/K_0 = \pm (\rho'^2/r^2 \mp \rho'^2/b^2) \quad (41)$$

$$\sigma_{\theta}'/K_0 \quad (42)$$

$$\sigma_z'/K_0 = -2\nu\rho'^2/b^2 - E\varepsilon_z/K_0 \quad (43)$$

The continuity condition of σ_r' determines the relation between ρ' and ρ as given by

$$\begin{aligned} & 1 - \rho'^2/b^2 + 2(1-\eta\beta) \log \frac{\rho'}{a} + \eta\beta(\rho'^2/a^2 - 1) \\ &= \frac{1}{2} (1 - \rho^2/b^2) + (1-\eta\beta) \log(\rho/a) + \frac{1}{2} \eta\beta(\rho^2/a^2 - 1) \end{aligned} \quad (44)$$

The yield surface moves in translation during elastic-plastic unloading according to

$$\alpha_{\theta}' = -\alpha_r' = c\varepsilon_{\theta}'^P \quad (45)$$

where $\varepsilon_{\theta}'^P$ is given by Eq. (37).

V. NUMERICAL RESULTS AND DISCUSSIONS. Consider a closed-end thick-walled cylinder with the following parameters: $b/a = 3$, $\nu = 0.3$, and $E/K_0 = 200$. The numerical results for the displacements, strains, and stresses during elastic-plastic loading and unloading are calculated. Figure 1 shows the relationship between the internal pressure factor (p/K_0) and the dimensionless elastic-plastic interface (ρ/a) for various values of the hardening parameter, $m = 0, 0.05, 0.1$, and 0.2 . The displacements at the inside and outside boundaries of the tube (U_a and U_b) are shown in Figure 2 as functions of the elastic-plastic interface for $m = 0.1$. The solid and dotted curves represent the displacements during loading and after unloading, respectively. Figure 3 shows the distribution of hoop stresses (σ_{θ}) during loading for $\rho/a = 1.0, 1.5, 2.0, 2.5$, and 3.0 . After complete unloading from different stages of loading, the corresponding residual hoop stresses (σ_{θ}'') are shown in Figure 4. Reverse yielding occurs in a strain-hardening tube with $m = 0.1$ only when the plastic portion (ρ) is larger than $1.652a$. In order to show the effect of hardening parameters (m) on the residual stress distribution, the numerical results are presented in Figure 5 for $m = 0, 0.05$, and 0.10 . As can be seen in the figure, larger values of hardening parameter (m) tend to reduce the beneficial residual hoop stress at the bore.

All the results presented in Figures 1 through 5 are based on the kinematic theory. We have also calculated the results based on the isotropic hardening theory [2]. Figure 6 shows a comparison of two hardening rules for the residual hoop stresses in a closed-end thick-walled cylinder. The dotted

curves represent isotropic hardening model with no Bauschinger effect. According to this model, there is no reverse yielding. The solid curves show the Bauschinger effect and reverse yielding occurs in both cases, $\rho'/a = 1.098, 1.336$, for $\rho/a = 2.0, 3.0$, respectively. The residual hoop stresses at the bore are $\sigma_\theta''/K_0 = -0.729, -0.327$ for $\rho/a = 2.0, 3.0$, respectively. According to isotropic hardening rule, those values of σ_θ''/K_0 should be $-1.060, -1.318$ for $\rho/a = 2.0, 3.0$, respectively. These numerical results indicate that the effect of hardening rules on the residual hoop stresses is quite significant, especially near the bore. Many plasticity theories for reverse yielding have been proposed and reviewed [9], and many computer programs have been developed [10]. It is believed that the numerical results based on other theories will fall within the limits obtained by using the kinematic and isotropic hardening models.

REFERENCES

1. Hill, R., Mathematical Theory of Plasticity, Oxford University Press, 1950.
2. Bland, D. R., "Elastoplastic Thick-Walled Tubes of Work-Hardening Materials Subject to Internal and External Pressures and Temperature Gradients," Journal of Mechanics and Physics of Solids, Vol. 4, 1956, pp. 209-229.
3. Chen, P. C. T., "The Finite Element Analysis of Elastic-Plastic Thick-Walled Tubes," Proceedings of Army Symposium on Solid Mechanics, The Role of Mechanics in Design-Ballistic Problems, 1972, pp. 243-253.
4. Chen, P. C. T., "Elastic-Plastic Analysis of a Radially Stressed Annular Plate," Journal of Applied Mechanics, Vol. 44, 1977, pp. 167-169.
5. Chen, P. C. T. and O'Hara, G. P., "Finite Element Solutions of Pressurized Thick Tubes," Proceedings of the Fifth Engineering Mechanics Division Specialty Conference, ASCE, 1984, pp. 1137-1140.
6. Prager, W., "The Theory of Plasticity: A Survey of Recent Achievements," Proceedings of the Institution of Mechanical Engineers, Vol. 169, 1955, pp. 41-57.
7. Ziegler, H., "A Modification of Prager's Hardening Rule," Q. Appl. Math., Vol. 17, 1959, pp. 55-65.
8. Bathe, K. J., "ADINA User's Manual," Report AE-81-1, ADINA Engineering Inc., Watertown, MA, 1981.
9. Armen, H., "Plasticity in General Software," Workshop on Inelastic Constitutive Equations for Metals, (E. Krempl, C. H. Wells, and Z. Zudans, Eds.) Rensselaer Polytechnic Institute, Troy, NY, 1975, pp. 56-78.
10. Noor, A. K., "Survey of Computer Programs for Solution of Nonlinear Structural and Solid Mechanics Problem," Computer and Structures, Vol. 13, 1981, pp. 425-465.

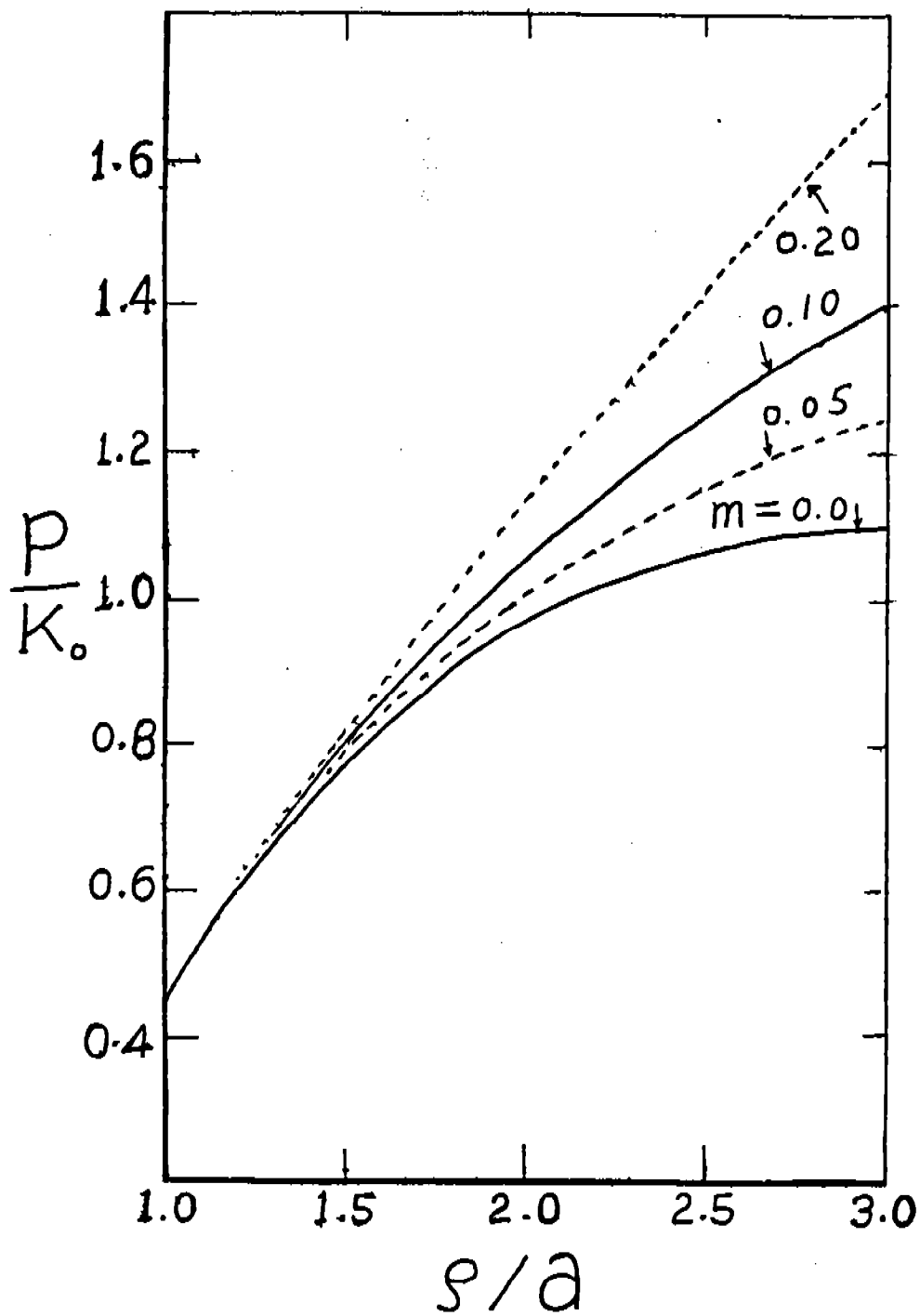


Figure 1. Effect of hardening on the relation between internal pressure and elastic-plastic interface.

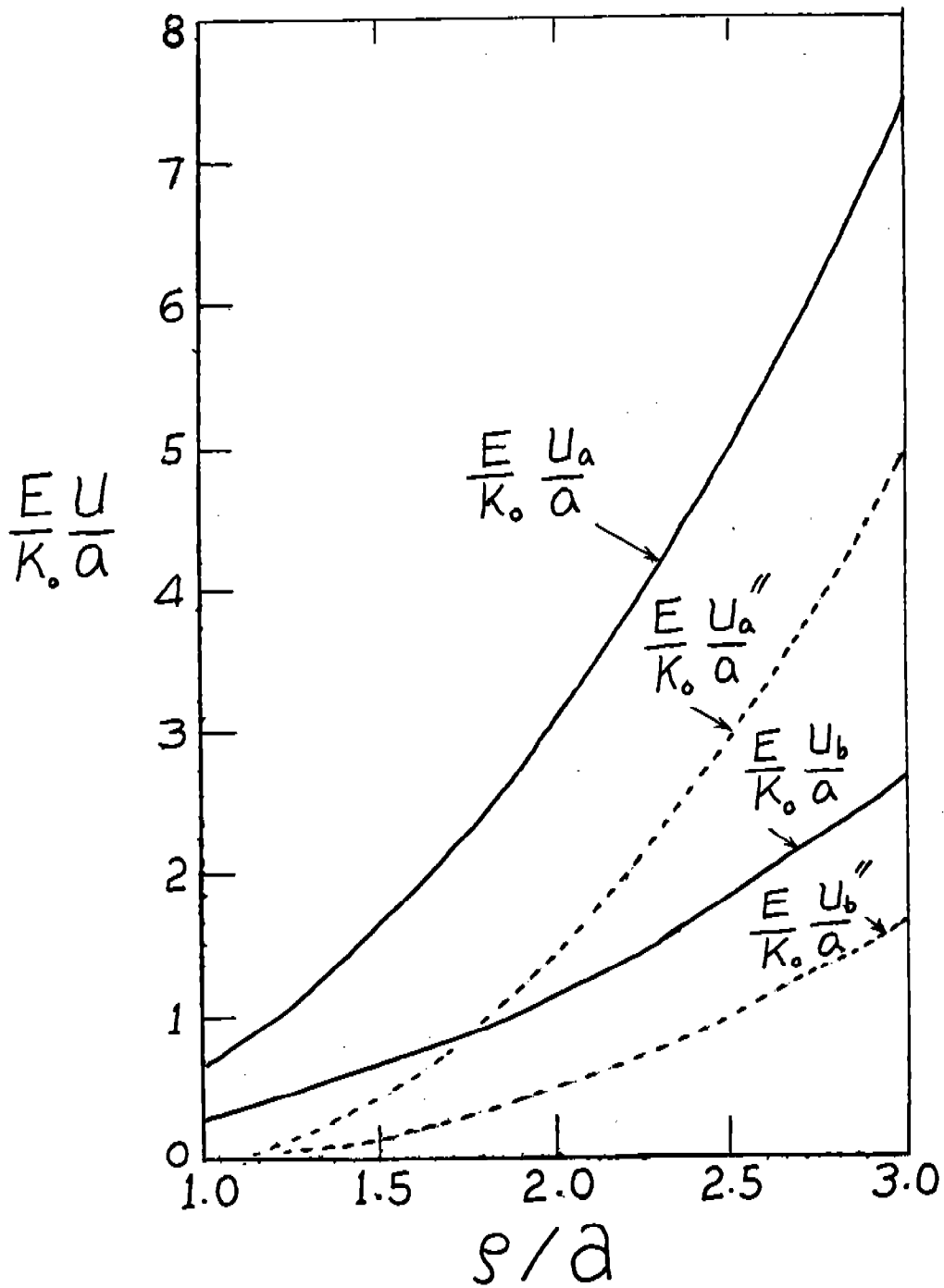


Figure 2. Boundary displacement during loading and after loading.

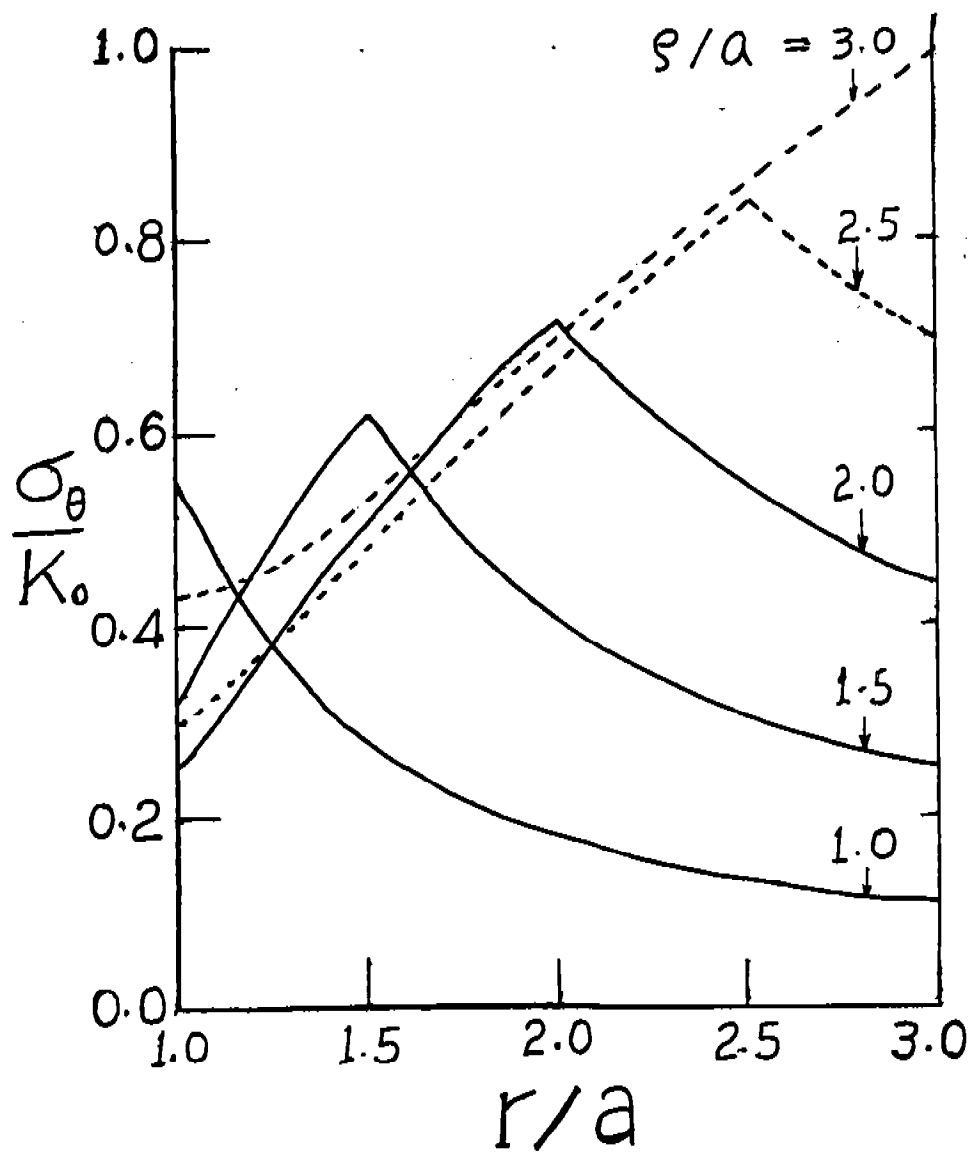


Figure 3. Distribution of hoop stresses during loading.

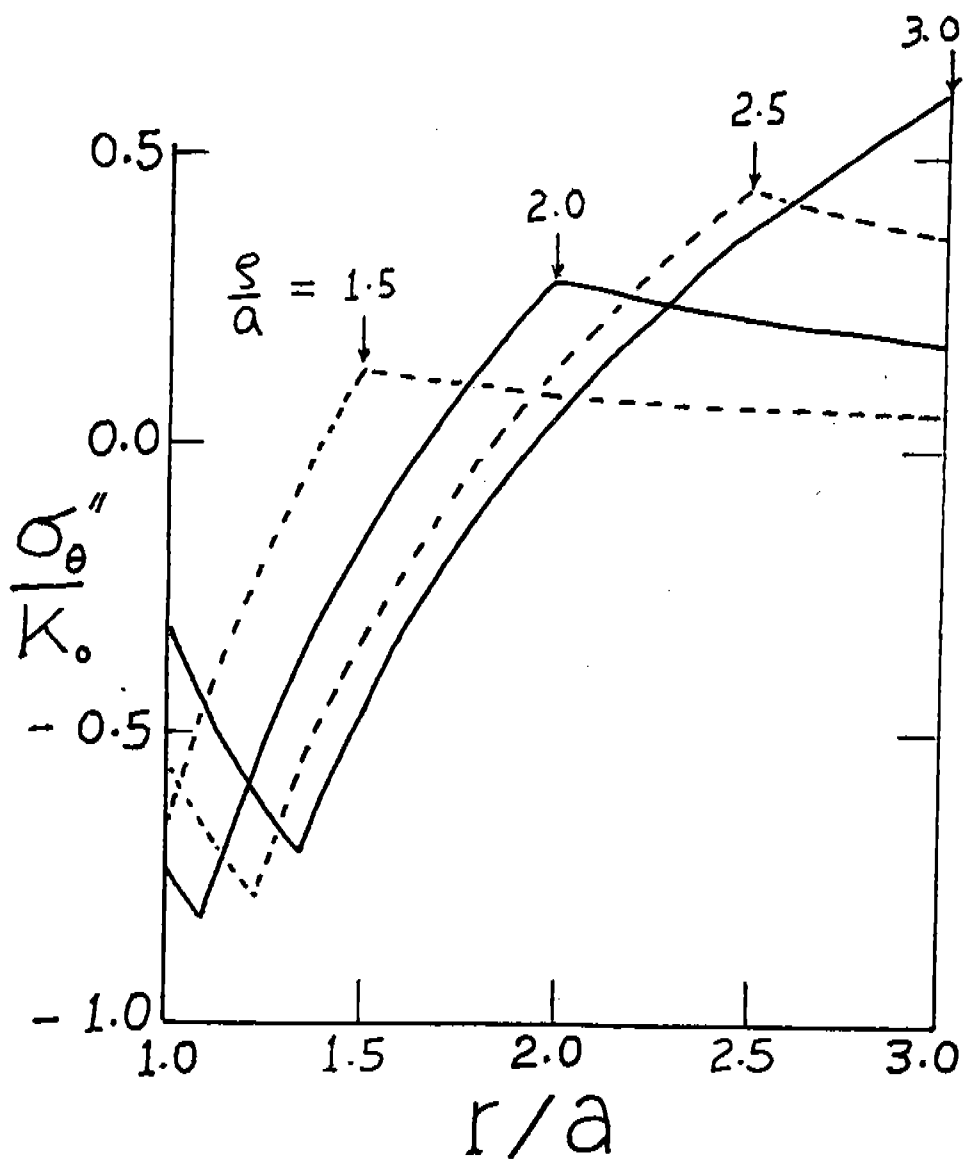


Figure 4. Distribution of residual hoop stresses.

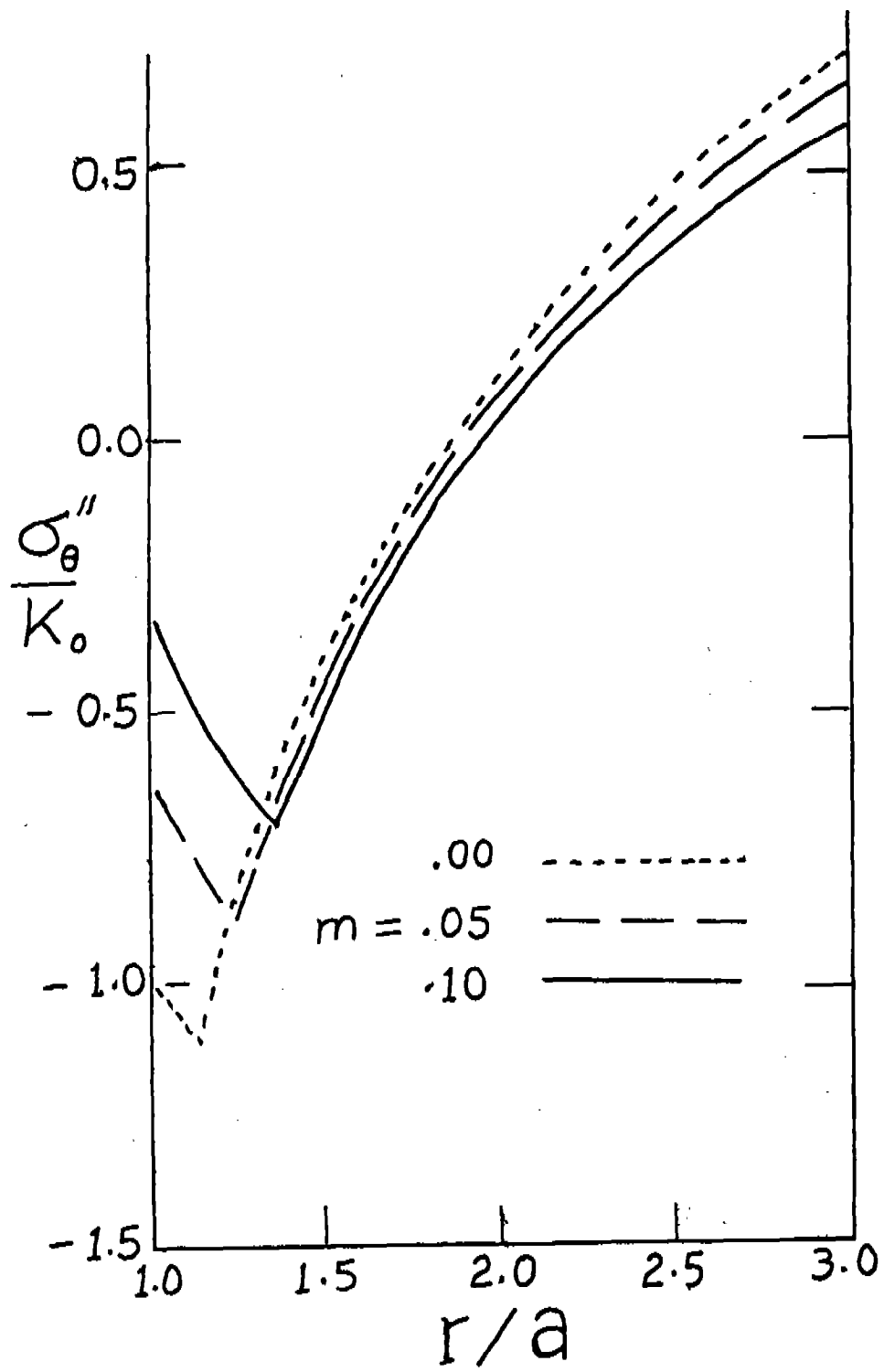


Figure 5. Effect of hardening parameter on residual stress distribution.

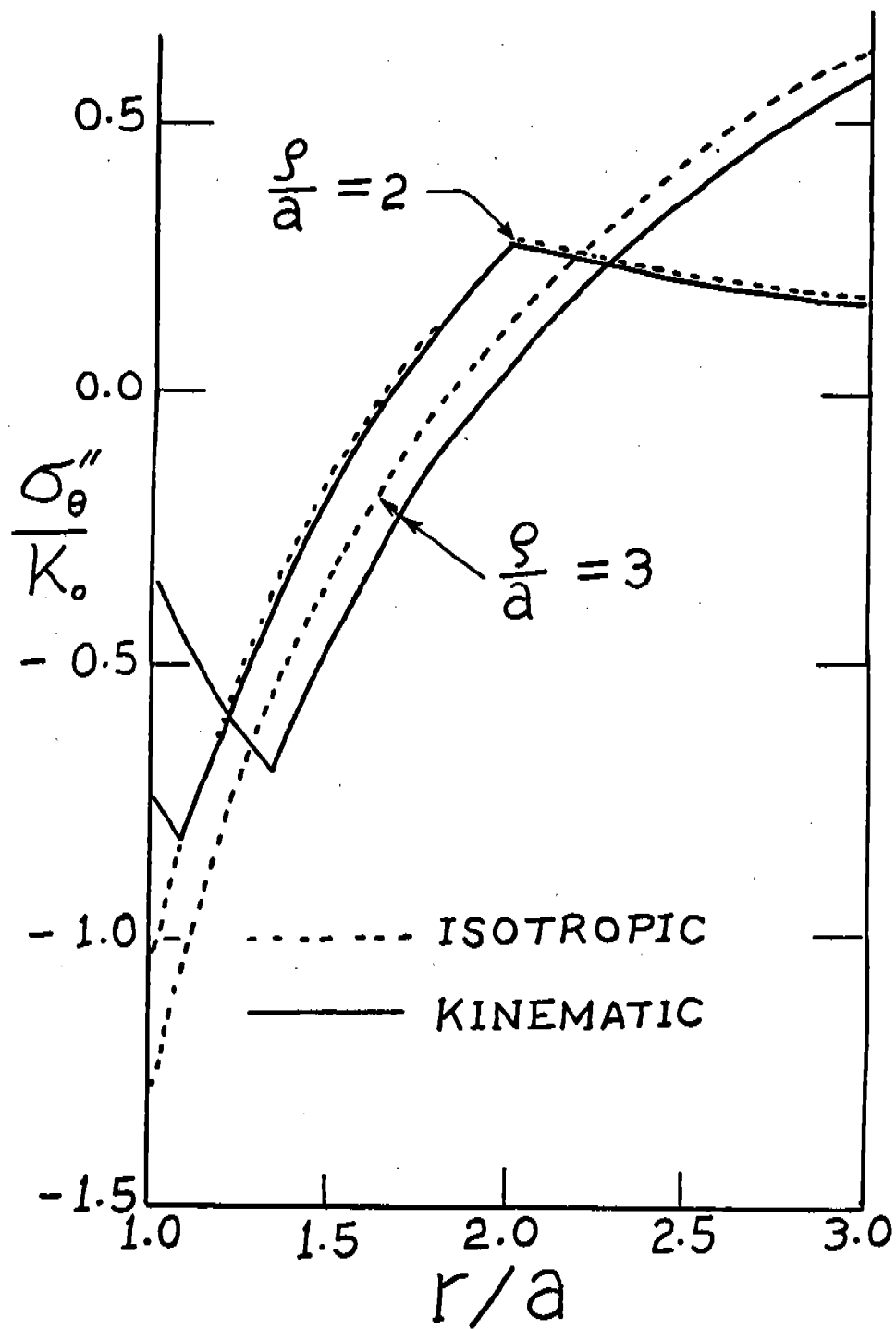


Figure 6. Effect of hardening rules on residual stress distribution.

A SIMPLIFIED ORTHOTROPIC FORMULATION
OF THE
VISCOPLASTICITY THEORY BASED ON OVERSTRESS

M. Sutcu and E. Krempl
Department of Mechanical Engineering,
Aeronautical Engineering & Mechanics
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

ABSTRACT

Representation theorems for tensor functions together with the uniaxial formulation of the theory of viscoplasticity based on overstress are used to present a fully invariant three-dimensional, small-strain, nonlinear orthotropic theory. To facilitate identification of the material functions, "elastic anisotropy" and "inelastic anisotropy" are related by a scalar function.

In each preferred direction the stress is composed of time (rate)-independent (or plastic) and viscous (or rate-dependent) contributions. It is demonstrated that these contributions can vary in the different preferred directions. It is possible to have almost linear elastic response in one direction while the other two directions exhibit inelastic deformation. While normally the stress-strain curve with the highest elastic modulus will also exhibit the highest stress in the plastic range, this tendency can be reversed by a judicious choice of the invariants.

The above properties are demonstrated by numerical experiments which include monotonic and cyclic loading at constant strain rate as well as relaxation tests.

1. INTRODUCTION

Recently the modeling of inelastic behavior of metals has been changing from predominantly yield surface oriented approaches to others, which use different concepts such as the endochronic time [1] for the time-independent case. For rate-dependence, the unified* theories [2-5] were developed which, like the endochronic theory, do not use the formalism associated with yield surface plasticity. These theories are proposed for the uniaxial case and for isotropy when three-dimensional formulations are considered.

The viscoplasticity theory based on overstress is of the unified type and considers rate-dependence fundamental [6,7]. It was previously developed as a deformation theory [8] and is now extended to cyclic neutral behavior [6,7]. As in the case of the unified theories, only an isotropic formulation has been proposed so far.

When considering anisotropy in the inelastic deformation of metals, it is necessary to distinguish between deformation induced anisotropy and natural anisotropy. For the former, an initially isotropic material may become permanently anisotropic after inelastic(plastic) deformation. The latter is essentially present in single crystals, metal matrix composites and directionally solidified alloys. Frequently the anisotropy induced by a kinematic variable (backstress) is also considered separately.

Within the context of the theory of viscoplasticity based on overstress, an orthotropic formulation is introduced in this paper. The aim of the formulation is not to provide the most general, tensorially invariant representation. Rather, a simplified version is developed which captures key anisotropic phenomena and renders the identification of real material

*The term "unified" is used to indicate that creep and plastic strains are not considered separately; all inelastic deformation is rate dependent and is modeled through the inelastic strain rate.

properties through testing as simple as possible. To this end it is assumed that elastic anisotropy and inelastic anisotropy are simply related. The deformation induced anisotropy is not included but anisotropy due to the development of the kinematic variable (backstress, equilibrium stress) is part of the formulation.

In the present analysis, tensor function representations and the uniaxial formulation of the viscoplasticity theory based on overstress [6] are used to synthesize an orthotropic representation using anisotropic elasticity theory as a guide. Tensor function representation theorems are very useful and, to the knowledge of the authors, have not been employed in the derivation of anisotropic viscoplasticity theories.

After some introductory remarks the evolution equations for the inelastic strain rate and for the equilibrium stress are given in chapter 2. To facilitate identification it is assumed that elastic anisotropy and inelastic anisotropy are the same, i.e. the respective "fourth-order material tensors" are related by a scalar multiple. Inelastic incompressibility is not considered in this paper but has been formulated elsewhere [9,10]. In chapter 3 the equations are specialized for various simple tests which are important in the identification procedure. It is demonstrated that the theory is capable of reproducing elastic response in one direction while the other two directions exhibit inelastic behavior. The capabilities of the proposed theory are illustrated by numerical experiments (numerical integration of the coupled nonlinear differential equation system using hypothetical but realistic material properties).

2. EVOLUTION EQUATIONS

2.1 Introductory Remarks

Anisotropy has been considered within the context of time-independent yield surface plasticity, e.g. [11,12] to cite just a few, and for viscoplasticity [13-16]. The present approach differs from the others by the use of representation theorems for isotropic tensor functions using "geometric" tensors and the usual deformation variables as arguments. To this end, the constitutive relations for orthotropic materials depend on three symmetric geometric tensors m , n , ℓ , which are formed as

$$m_{ij} = M_i M_j \quad (1)$$

$$n_{ij} = N_i N_j \quad (2)$$

$$\ell_{ij} = L_i L_j \quad (3)$$

as well as the usual mechanical variables. The unit vectors M_i , N_i , and L_i are mutually orthogonal to each other and parallel to the intersections of the three orthogonal symmetry planes which define the state of orthotropy.

The infinitesimal total strain rate is the sum of elastic and inelastic parts

$$\dot{\epsilon}_{ij} = \dot{\epsilon}_{ij}^e + \dot{\epsilon}_{ij}^{in} \quad (4)$$

The elastic strain rate is governed by the hypoelastic law

$$\dot{\epsilon}_{ij}^e = S_{ijpq} \dot{\sigma}_{pq} \quad (5)$$

The inverse of (5) is

$$\dot{\sigma}_{ij} = C_{ijpq} \dot{\epsilon}_{pq}^e \quad (6)$$

Representations of S and C for orthotropy are given in the Appendix.

Using (4), Eq.(6) can be rewritten as

$$\dot{\sigma}_{ij} = C_{ijpq} (\dot{\epsilon}_{pq} - \dot{\epsilon}_{pq}^{in}) \quad (7)$$

2.2 Evolution of Inelastic Strain

The inelastic strain rate $\dot{\epsilon}_{pq}^{in}$ is assumed to depend on a quantity called overstress [6-8] and denoted by X_{ij}

$$\dot{\epsilon}_{ij}^{in} = F_{ij}[X_{pq}, m_{pq}, n_{pq}, \ell_{pq}] \quad (8)$$

where

$$X_{pq} = \sigma_{pq} - g_{pq} \quad (9)$$

The quantity g_{pq} is the equilibrium stress for which an evolution law is needed. The function F_{ij} is an isotropic tensor valued function of the arguments enclosed by square brackets.

The complete representation of F_{ij} is available from [17,18]

$$\begin{aligned} \dot{\epsilon}^{in} = & k_1 \underline{\underline{m}} + k_2 \underline{\underline{n}} + k_3 \underline{\underline{\ell}} + k_4 (\underline{\underline{m}} \underline{\underline{X}} + \underline{\underline{X}} \underline{\underline{m}}) \\ & + k_5 (\underline{\underline{n}} \underline{\underline{X}} + \underline{\underline{X}} \underline{\underline{n}}) + k_6 (\underline{\underline{\ell}} \underline{\underline{X}} + \underline{\underline{X}} \underline{\underline{\ell}}) \\ & + k_7 \underline{\underline{X}}^2 \end{aligned} \quad (10)$$

where the scalar functions k_i depend on seven invariants

$$\begin{aligned} & \text{tr } \underline{\underline{m}} \underline{\underline{X}}, \quad \text{tr } \underline{\underline{n}} \underline{\underline{X}}, \quad \text{tr } \underline{\underline{\ell}} \underline{\underline{X}} \\ & \text{tr } \underline{\underline{m}} \underline{\underline{X}}^2, \quad \text{tr } \underline{\underline{n}} \underline{\underline{X}}^2, \quad \text{tr } \underline{\underline{\ell}} \underline{\underline{X}}^2, \quad \text{tr } \underline{\underline{X}}^3 \end{aligned} \quad (11)$$

By choosing

$$k_7 = 0 \quad (12)$$

the quadratic term $\underline{\underline{X}}^2$ is eliminated in the present study in accordance with the uniaxial overstress model [6-8].

Retention of this term offers the possibility of modeling second-order effects within the context of finite deformation which will not be explored here.

Thus the complete representation of (10) can be given as

$$\dot{\epsilon}_{ij}^{\text{in}} = k_{ijpq} \dot{X}_{pq} \quad (13)$$

where the components of the fourth-order tensor k are functions of the scalar invariants and the geometric tensors given in (11) and (1)-(3), respectively.

In order to reduce the complexity of the fourth-order tensor k further, we will introduce the concept of "anisotropy ratios." These ratios can also be defined for elasticity. As an example, the ratio of Young's moduli in two different directions in the material is an "anisotropic ratio." Thus, (10) becomes

$$\dot{\epsilon}_{ij}^{\text{in}} = k r_{ijpq} \dot{X}_{pq} \quad (14)$$

The components of the dimensionless quantity r are called "inelastic anisotropy ratios." In general they may be functions but will be constants in this paper. The representation of r for orthotropy is given in (A-10).

The quantity k has dimensions of (time-stress)⁻¹ and is a scalar function of the set of invariants given in (11). This function is simply related to "the viscosity function" of the isotropic or uniaxial theory [6-8]. It governs the strain-rate dependence of the uniaxial stress-strain curves in different material directions.

For practical reasons a single argument denoted by Γ is proposed for the viscosity function k

$$\Gamma^2 = \text{tr}(\underline{r} \underline{\dot{X}})(\underline{r} \underline{\dot{X}}) \quad (15)$$

It is possible to use a different \bar{r} in (15) than the one given in (14).

2.3 Evolution of the Equilibrium Stress

An evolution law for the equilibrium stress g_{ij} can be developed from the representation theorems [17,18]. In generalizing the uniaxial over-stress model, the rate of equilibrium stress is an isotropic tensor function of six symmetric tensors

$$\dot{g}_{ij} = H_{ij} [\dot{\epsilon}_{pq}^{in}, \dot{\epsilon}_{pq}, \bar{g}_{pq}, m_{pq}, n_{pq}, l_{pq}] \quad (16)$$

where

$$\bar{g}_{ij} = g_{ij} - f_{ij} \quad (17)$$

and f_{ij} is the generalization of the f function of [6]. Its derivative represents the final slope of the stress-strain relation in the plastic range.

The complete representation of (16) is too general and lengthy for present purposes. Instead, the forms proposed in [10] will be considered (see Eqs.(18) and (20) of [10]). A special case which retains sufficient generality for the present is

$$\dot{g}_{ij} = \psi r_{ijpq}^{-1} \dot{\gamma}_{pq} \quad (18)$$

The scalar ψ is a decreasing function of overstress X and is initially slightly less than Young's modulus in one of the preferred directions, e.g.

$$\psi[0] = \bar{E}_1 < E_1^\dagger \quad (19)$$

The anisotropy ratio \bar{r}^{-1} is described after (A-10).

[†]If $\bar{E}_1 = E_1$, then (7), (14) and (18) represent only linear elastic behavior. Any positive $\bar{E}_1 < E_1$ results in inelastic behavior. Usually $\bar{E}_1 = 0.99 E_1$.

The argument $\dot{\gamma}_{pq}$ has the following form

$$\dot{\gamma}_{pq} = \dot{\epsilon}_{pq} - \Gamma_g^2 (1 - \phi) \dot{\epsilon}_{pq}^{\text{in}} \quad (20)$$

where Γ_g and ϕ are dimensionless scalar functions of invariants of the modified equilibrium stress \bar{g} and overstress X , respectively.

The scalar function Γ_g^2 in (20) governs the anisotropy of the equilibrium stress and is given as

$$\Gamma_g^2 = \text{tr}(\bar{r}_{\approx} \bar{g})(\bar{r}_{\approx} \bar{g}) / A^2 \quad (21)$$

where \bar{g} is defined in (17). Depending upon modeling requirements, a quantity \bar{r}_{\approx} , which is different than that of (14) or (15), can be used in (21). The constant A^2 is the asymptotic value of $\text{tr}(\bar{r}_{\approx} \bar{g})(\bar{r}_{\approx} \bar{g})$ in a constant strain-rate test.

The function f_{ij} defined in (17) is assumed to be

$$f_{ij} = P_1 r_{ijpq}^{-1} \epsilon_{pq} \quad (22)$$

The constant P_1 represents the slope in the inelastic range and is reached asymptotically.

The scalar function ϕ in (20) is selected as

$$\phi[\Gamma] = \frac{P_1}{\psi[\Gamma]} \quad (23)$$

where P_1 and ψ are used in (22) and (18), respectively.

3. UNIAXIAL TESTS IN PREFERRED DIRECTIONS FOR CONSTANT ANISOTROPIC RATIOS

If we assume that anisotropic ratios remain the same for both elastic and inelastic deformations, then the general formulation of the previous section simplifies considerably. In this case

$$\underline{\underline{r}} = E_1 \underline{\underline{S}} , \quad (24)$$

$$\underline{\underline{r}}^{-1} = \underline{\underline{C}}/E_1 , \quad (25)$$

where the quantities $\underline{\underline{S}}$, $\underline{\underline{C}}$ and $\underline{\underline{r}}$ were introduced in (5), (6) and (14). The explicit forms of $\underline{\underline{S}}$ and $\underline{\underline{C}}$ are given in the Appendix.

The stress- and strain-tensor components are assumed to be given in the preferred coordinate system represented by $\underline{\underline{M}}$, $\underline{\underline{N}}$ and $\underline{\underline{L}}$. The uniaxial tests are carried out in the preferred directions and then the geometric tensor components of (1) - (3) are

$$\begin{aligned} \text{all } m_{ij} &= 0 & \text{except } m_{11} &= 1 \\ \text{all } n_{ij} &= 0 & \text{except } n_{22} &= 1 \\ \text{all } l_{ij} &= 0 & \text{except } l_{33} &= 1 . \end{aligned} \quad (26)$$

3.1 Uniaxial Tension in the z-Direction

In a uniaxial tension test in the preferred z or 3 direction all the components of the stress tensor are zero, except σ_z which corresponds to the applied load. The lateral components of the equilibrium stress tensor are zero. The shear components vanish due to material symmetry.

The evolution law for σ_z is obtained from (7) by using the quantity $\underline{\underline{C}}$ given in the Appendix

$$\dot{\sigma}_z = E_3 (\dot{\epsilon}_z - \dot{\epsilon}_z^{\text{in}}) \quad (27)$$

where

$$\dot{\epsilon}_z^{\text{in}} = r_2 (\sigma_z - g_z) k . \quad (28)$$

The strain rate $\dot{\epsilon}_z = \text{constant}$ is the input parameter.

Since the formulation corresponds to a constant Poisson's ratio, the lateral strain rates are also constant

$$\dot{\epsilon}_x = -v_{31} \dot{\epsilon}_z \quad (29)$$

$$\dot{\epsilon}_y = -v_{32} \dot{\epsilon}_z \quad (30)$$

The lateral inelastic strain rates are

$$\dot{\epsilon}_x^{in} = -r_{31} \dot{\epsilon}_z^{in} \quad (31)$$

$$\dot{\epsilon}_y^{in} = -r_{32} \dot{\epsilon}_z^{in} \quad (32)$$

The argument Γ of the function k is determined from (15) using (28), (31) and (32)

$$\Gamma_1^2 = (r_{31}^2 + r_{32}^2 + 1)r_2^2(\sigma_z - g_z)^2 \quad (33)$$

The only nonzero component of the equilibrium stress g_z has the following evolution law which is obtained from (18) using (25),

$$\dot{g}_z = \frac{\psi}{r_2} \left(\dot{\epsilon}_z - \Gamma_g (1 - \phi) \dot{\epsilon}_z^{in} \right) \quad (34)$$

where Γ_g is obtained from (21) as

$$\Gamma_{g_1}^2 = (r_{31}^2 + r_{32}^2 + 1)r_2^2(g_z - f_z)^2/A^2 \quad (35)$$

and ϕ is given by (23). Note that all components of f_{ij} are zero except

$$f_z = P_1 \frac{E_3}{E_1} \epsilon_z \quad (36)$$

as can be seen from (22).

For simplicity let us assume that f_{ij} is zero, and set $P_1 = 0$.

The asymptotic solutions of (27) and (34) will occur when

$$\dot{\epsilon}_z = \dot{\epsilon}_z^{in} = r_2(\sigma_z - g_z)k[\Gamma_1] \quad (37)$$

and

$$\frac{r^2}{g_1} = 1, \quad (38a)$$

or, correspondingly

$$g_z = \frac{A}{r_2(r_{31}^2 + r_{32}^2 + 1)^{1/2}} \quad (38b)$$

which render $\dot{\sigma}_z = \dot{g}_z = 0$.

Notice that the final value of the equilibrium stress g_z is a constant which is independent of the input strain rate $\dot{\epsilon}_z$.

Equation (37) is a nonlinear algebraic expression of the over-stress $\sigma_z - g_z$. For different input strain rates, different asymptotic overstress values are obtained.

3.2 Uniaxial Tension in the x-Direction

Similar procedure is applied to the case of constant strain-rate loading in the x-direction. We will compare the asymptotic properties of this case to that of the previous case.

The evolution of σ_x is governed by

$$\dot{\sigma}_x = E_1(\dot{\epsilon}_x - \dot{\epsilon}_x^{in}) \quad (39)$$

where

$$\dot{\epsilon}_x^{in} = (\sigma_x - g_x)k \quad (40)$$

The lateral total and inelastic strain rates are governed by

$$\dot{\epsilon}_y = -\nu_{21} \frac{E_1}{E_2} \dot{\epsilon}_x \quad (41)$$

$$\dot{\epsilon}_z = -\nu_{31} \frac{E_1}{E_3} \dot{\epsilon}_x \quad (42)$$

$$\dot{\epsilon}_y^{in} = -r_{21} r_1 \dot{\epsilon}_x^{in} \quad (43)$$

$$\dot{\epsilon}_z^{in} = -r_{31} r_2 \dot{\epsilon}_x^{in} \quad (44)$$

The argument Γ of the function k is

$$\Gamma_2^2 = \left((r_{21} r_1)^2 + (r_{31} r_2)^2 + 1 \right) (\sigma_x - g_x)^2 \quad (45)$$

The nonzero component of the equilibrium stress g_x is governed by

$$\dot{g}_x = \psi \left(\dot{\epsilon}_x - \Gamma_g (1 - \phi) \dot{\epsilon}_x^{\text{in}} \right) \quad (46)$$

where functions ψ and ϕ are the same as in the previous case, and Γ_g is

$$\Gamma_{g_2} = \left((r_{21} r_1)^2 + (r_{31} r_2)^2 + 1 \right) (g_x - f_x)^2 / A^2 \quad (47)$$

The nonzero component f_x of f_{ij} is given by

$$f_x = P_1 \epsilon_x \quad (48)$$

The asymptotic solutions of (39) and (46), when $P_1 = 0$ and $\phi = 0$, will occur when

$$\dot{\epsilon}_x = \dot{\epsilon}_x^{\text{in}} = (\sigma_x - g_x) k[\Gamma_2] \quad (49)$$

and

$$g_x = \frac{A}{\left((r_{21} r_1)^2 + (r_{31} r_2)^2 + 1 \right)^{1/2}} \quad (50)$$

Thus the final value of the equilibrium stress g_x is the constant on the right-hand side of (50).

The spacing of the equilibrium stresses g_x and g_z of these two uniaxial tests is governed by

$$\frac{g_x}{g_z} = r_2 \frac{(r_{31}^2 + r_{32}^2 + 1)^{1/2}}{\left((r_{21} r_1)^2 + (r_{31} r_2)^2 + 1 \right)^{1/2}} \quad (51)$$

which is obtained from (50) and (38b). Since the equilibrium stress is a measure of time-independent change in the material, the ratio in (51) can be called "plastic anisotropy."

The strain-rate spacing of the uniaxial test curves in the x-direction is governed by the nonlinear algebraic equation (49). Different spacings predicted in x- and z-directions by (49) and (37) for the same input strain rate can be termed "viscous anisotropy." The ratio of the spacings is governed by

$$\frac{\sigma_x - g_x}{\sigma_z - g_z} = \frac{r_2 k[\Gamma_1]}{k[\Gamma_2]} \quad (52)$$

where $\dot{\epsilon}_z = \dot{\epsilon}_x$ and Γ_1 and Γ_2 are given by (33) and (45), respectively.

4. DISCUSSION

4.1 General

The fully invariant three-dimensional formulation is a simplified version of the theory presented in [9]. Here it is assumed that the elastic anisotropy ratio $E_1 \approx S$ (see (A.10)) is equal to \bar{r} appearing in (14), (18) and (22) which govern the evolution of the inelastic strain rate. This choice simplifies the identification of the material functions to be discussed shortly. As a consequence of this choice, constant Poisson's ratios are found in the uniaxial tests; see (29)-(32) and (41)-(44). This restriction, which is judged to be acceptable for the present purpose, is removed in [9] where a formulation with variable Poisson's ratio is given.

4.2 Identification

Rate dependence of the uniaxial stress-strain curves in a particular direction in the material are governed by the function k^\dagger in a nonlinear manner,

[†]The function k in this paper is equivalent to $(1/Ek)$ in previous papers.

see (37) or (49), and the invariant Γ . They control viscous anisotropy.

In principle the identification procedure is as follows. Uniaxial tests in a certain preferred direction, say x , are performed at piecewise constant strain rates during one test. Then a k function is introduced which can reproduce the same strain-rate spacing as found in the experiment. Thus the relaxation and creep properties in the x -direction are also determined, see [6].

The strain-rate spacing (and correspondingly the creep and the relaxation properties) in some other direction, say z , can be adjusted through the quantity \tilde{r} which is used to calculate Γ in (15). It is not necessarily the same as that of (14), (18) and (22), although this is assumed in the body of the paper. (An exception is made in the numerical examples to show the versatility of the approach.) An analytical expression is given in (52) for the ratio of the uniaxial overstress values in the x - and z -directions. This ratio depends on r_2 of \tilde{r} in (14), and the function k evaluated at two different arguments Γ_1 and Γ_2 , which are given by (33) and (45) for x - and z -directions, respectively. In such a way, viscous anisotropy is reproduced.

The plastic (rate-independent) anisotropy is governed by the asymptotic spacing of the uniaxial equilibrium stress curves and controlled by the invariant Γ_g . As in the case of Γ , quantity \tilde{r} , which is used to calculate Γ_g in (21), does not have to be the same as that of (14) or (15). The asymptotic g values for uniaxial tests in x - and z -directions and their ratio are given by (35b), (50) and (51) in terms of the particular \tilde{r} in (21). They can be used for the adjustment of material data.

4.3 Illustration of the Theory by Numerical Experiments

It was mentioned that the proposed simplified approach offers flexibilities by choosing the anisotropy ratios in the invariants Γ and Γ_g

in different ways. Specifically three cases can be identified.

- (i) The quantities $\underline{\underline{r}}$ in (14), (15) and (21) are all equal to each other which are, in turn, set equal to $E_1 \underline{\underline{S}}$.
- (ii) The quantity $\underline{\underline{r}}$ for the inelastic strain rate in (14) is set equal to $E_1 \underline{\underline{S}}$. The quantities in (15) and (21) are equal but different than $E_1 \underline{\underline{S}}$.
- (iii) The quantity $\underline{\underline{r}}$ in (14) is set equal to $E_1 \underline{\underline{S}}$. The other quantities in (15) and (21) are chosen independently.

In case (i) the inelastic properties, including viscous and plastic anisotropy, are nonlinearly governed by the elastic properties.

This behavior is illustrated by numerical experiments (the numerical integration of the coupled system of nonlinear differential equations) using the material data listed in Table 1. It should be noted that the material constants were chosen so that transverse isotropy, a special case of orthotropy is modeled. The preferred direction is the z- (or 3-) direction. As a consequence the behaviors in the x- and y-directions are identical.

Figure 1 shows the results of a tensile test with a strain rate of $\dot{\epsilon} = 10^{-4} \text{ s}^{-1}$. Both the components of σ and g in the z- and x-directions are plotted. It is seen that $\sigma_x > \sigma_z$ and $g_x > g_z$ at all times corresponding to $(E_x = E_1) > (E_z = E_3)$. At the end of the graph the asymptotic values are approximately reached and correspond to the theoretical predictions. (It should be noted that in Fig. 1 and in subsequent figures P_1 is chosen differently from zero resulting in a positive final slope of the σ and g curves.)

The relaxation behavior and the hysteresis loops predicted by the theory are plotted in Figs. 2 and 3, respectively.

The different amounts of stress reduction during the 240 s relaxation time (AB vs. A'B' in Fig. 2) are obvious. In the cyclic case a considerable amount of Bauschinger effect is evident. It is also of interest that the

hysteresis loops are not quite closed indicative that the asymptotic values are not yet reached at 1.2%. (The theory presented herein represents cyclic neutral behavior, see [6,7].) In regions where the σ - and g -curves coincide elastic behavior is represented. It can be observed that the elastic regions in the two directions are different.

In cases (ii) and (iii) the viscous and the plastic anisotropy become uncoupled from the elastic properties. It is now possible to adjust the ratio of the uniaxial overstress values and/or the asymptotic spacing of the uniaxial g -curves independently.

However, it should be noted that the ratios of the lateral strains to the uniaxial strains remain fixed and are governed by the elastic Poisson's ratios.

A numerical example is given for case (ii) in Fig. 4. The material data is the same as that of Figs. 1,2 and 3, except r_2 in (15) and (21) is set equal to zero. Thus the invariants Γ and Γ_g become zero, see (33) and (35).

As a consequence k in (28) is constant and equal to $k[0]$ which is typically a very small number. This choice renders the inelastic strain rate very small. Similarly the evolution of g_z is a straight line with modulus $\psi[0]/r_2$. This choice gives nearly elastic behavior in the z -direction without, of course, affecting the response in the x -direction. These properties are illustrated in Fig. 4

Another interesting example of case (ii) is the modeling of nearly elastic response under a hydrostatic state of stress

$$r_{21} = \frac{1}{2} \left(1 + \frac{1}{r_1} - \frac{r_2}{r_1} \right) \quad (53)$$

$$r_{31} = \frac{1}{2} \left(1 + \frac{1}{r_2} - \frac{r_1}{r_2} \right) \quad (54)$$

$$r_{32} = \frac{1}{2} \left(1 + \frac{r_1}{r_2} - \frac{1}{r_2} \right) \quad (55)$$

in (15) and (21), the invariants Γ and Γ_g will filter out all the hydrostatic components. However the trace of the inelastic strain rate in (14) is not zero.

It can be shown that, if a hydrostatic state of stress is applied, the equilibrium stress and consequently the overstress are also hydrostatic. The choice of (53) - (55) in (15) and (21) yields $\Gamma = \Gamma_g = 0$ in this case. Thus the behavior is again nearly elastic.

An example of case (iii) is given in Fig. 5. The material data is the same as that of Figs. 1, 2 and 3, except r_2 in Γ_g is set equal to 0.5 (note that $r_2 = E_1/E_3 = 50/35$ otherwise). The behavior in the x-direction is not affected by this choice; compare Fig. 5 to Fig. 1. The curves for uniaxial σ_z and g_z coincide with those of Fig. 1 in the elastic region. Although the Young's modulus in the z-direction E_3 is smaller than E_1 , the asymptotic value of the g_z -curve is higher than that of g_x due to the choice of $r_2 = 0.5$ in Γ_g ; see also (51). (Due to the choice of functions, the evolution for g_z has not reached the asymptotic value at the strain limit of the graph.) The amount of overstress in the z-direction is not affected by Γ_g ; compare Fig. 1 and Fig. 5.

4.4 Additional Remarks

The inelastic deformation is not volume preserving in the present formulation. However, a model with isochoric inelastic deformations is developed in [9] and will be presented elsewhere. The theory is intended for use with composites as well as "naturally" anisotropic metals. Since experimental confirmation of the isochoric nature of inelastic deformation of these materials seems to be absent, both formulations were developed in [9].

When initially loaded, the response of the present theory is symmetric with respect to the stress-strain origin; i.e., the stress-strain diagrams in tension and compression are congruent. When other than purely quadratic invariants are retained as arguments in the functions, e.g. in ψ , "strength differential effects" can be reproduced. In this case the stress-strain diagrams in tension and compression are not congruent. Examples of models of such behavior are given in [9].

Acknowledgment

The work was started under NASA Grant NAG3-262. Partial support was received from the U.S. Army Research Office Center of Excellence Contract DAAG 29-82-K-0093, Dr. Robert Singleton, technical monitor. The second author was supported by the National Science Foundation Solid Mechanics Program Grant MEA83-15967. The plotting programs were prepared by David Yao.

REFERENCES

- [1] K. C. Valanis, "Fundamental Consequences of a New Intrinsic Time Measure. Plasticity as a Limit of the Endochronic Theory," *Arch. Mech.*, 32, 171-191 (1980).
- [2] A. K. Miller, "An Inelastic Constitutive Model for Monotonic, Cyclic and Creep Deformation, Parts I and II," *Trans. ASME, J. Eng. Matls. and Tech.*, 98, 97-113 (1976).
- [3] E. W. Hart, "Constitutive Relations for the Nonelastic Deformation of Metals," *Trans. ASME, J. Eng. Matls. and Tech.*, 98, 193-201 (1976).
- [4] R. W. Rhode, and J. C. Swearingen, "Deformation Modeling Applied to Stress Relaxation of Four Solder Alloys," *Trans. ASME, J. Eng. Matls. and Tech.*, 102, 207-214 (1980).
- [5] S. R. Bodner, and Y. Partom, "Constitutive Equations for Elastic-Viscoplastic Strain-Hardening Materials," *Trans. ASME, J. Appl. Mech.*, 42, 385-389 (1975).
- [6] E. Krempl, J. J. McMahon, and D. Yao, "Viscoplasticity Based on Overstress with a Differential Growth Law for the Equilibrium Stress," 2nd Symp. on Nonlinear Constitutive Relations for High Temperature Applications, NASA, Cleveland, OH (1984), in press, *Mechanics of Materials*.
- [7] D. Yao, and E. Krempl, "Viscoplasticity Theory Based on Overstress. The Prediction of Monotonic and Cyclic Proportional and Nonproportional Loading Paths of an Aluminum Alloy," to appear in the *Int. J. of Plasticity*.
- [8] E. P. Cernocky, and E. Krempl, "A Theory of Viscoplasticity Based on Infinitesimal Total Strain," *Acta Mechanica*, 36, 263-289 (1980).
- [9] M. Sutcu, Ph.D. thesis, forthcoming (1985).
- [10] M. Sutcu, and E. Krempl, "A Transversely Isotropic Formulation of the Viscoplasticity Theory Based on Overstress," RPI Report MML 84-5, December (1984).
- [11] A. Baltov, and A. Sawczuk, "A Rule of Anisotropic Hardening," *Acta Mechanica*, Vol. I/2, pp. 81-92 (1965).
- [12] C. Shih, and D. Lee, "Further Developments in Anisotropic Plasticity," *ASME Journal of Engineering Materials and Technology*, Vol. 100, p. 294, (1978).
- [13] D. C. Stouffer, and S. R. Bodner, "A Constitutive Model for the Deformation Induced Anisotropic Plastic Flow of Metals," *Int. J. Eng. Sci.*, 17, 757-764 (1979).
- [14] D. N. Robinson, "Constitutive Relations for Anisotropic High Temperature Alloys," *NASA TM 83437*, Aug. 1983.
- [15] R. W. Young, "A Note on the Stouffer-Bodner Constitutive Model for Anisotropic Plastic Flow," *Lett. Appl. Eng. Sci.*, 18, 1091-1093 (1980).

- [16] L. Thomas Dame, "Anisotropic Constitutive Model for Nickel Base Single Crystal Alloys: Development and Finite Element Implementation," Ph.D. Thesis, University of Cincinnati (1985).
- [17] A.J.M. Spencer, in *Continuum Physics, Vol. 1 - Mathematics*, A. C. Eringen, Editor, Academic Press, New York (1971).
- [18] J. P. Boehler, "A Simple Derivation of Representations for Non-Polynomial Constitutive Equations in Some Cases of Anisotropy," *ZAMM*, 59, 157-167 (1979).
- [19] R. M. Jones, *Mechanics of Composite Materials*, McGraw-Hill Book Co., Washington, DC (1975).

Table 1

MATERIAL CONSTANTS AND FUNCTIONS USED

E_1	=	Modulus of Elasticity in the x-direction	=	50,000 MPa
E_2	=	" " " " " y-direction	=	E_1
E_3	=	" " " " " z-direction	=	35,000 MPa
ν_{21}	=	Poisson's ratio for the x- and y-directions	=	0.35
ν_{31}	=	" " " " " x- " z-directions	=	0.45
ν_{32}	=	" " " " " y- " z-directions	=	ν_{31}
r_1	=	E_1/E_2	=	1
r_2	=	E_1/E_3		
r_{21}	=	ν_{21}	Unless otherwise indicated	
r_{31}	=	ν_{31}		
r_{32}	=	ν_{32}		
P_1	=	500 MPa		
A^2	=	$22,500 \cdot ((r_{21} r_1)^2 + (r_{31} r_2)^2 + 1)$		$(\text{MPa})^2$
$k[\Gamma]$	=	$\frac{2.5 \times 10^{-6}}{E_1} \left(1 + \frac{\arg}{122.5}\right)^4$		$(\text{MPa sec})^{-1}$
$\psi[\Gamma]$	=	$E_1 \cdot (0.4 + 0.5 \exp(-6 \cdot 10^{-3} \arg))$		(MPa)
where				
\arg	=	$\frac{\Gamma^2}{((r_{21} r_1)^2 + (r_{31} r_2)^2 + 1)}$		$(\text{MPa})^2$

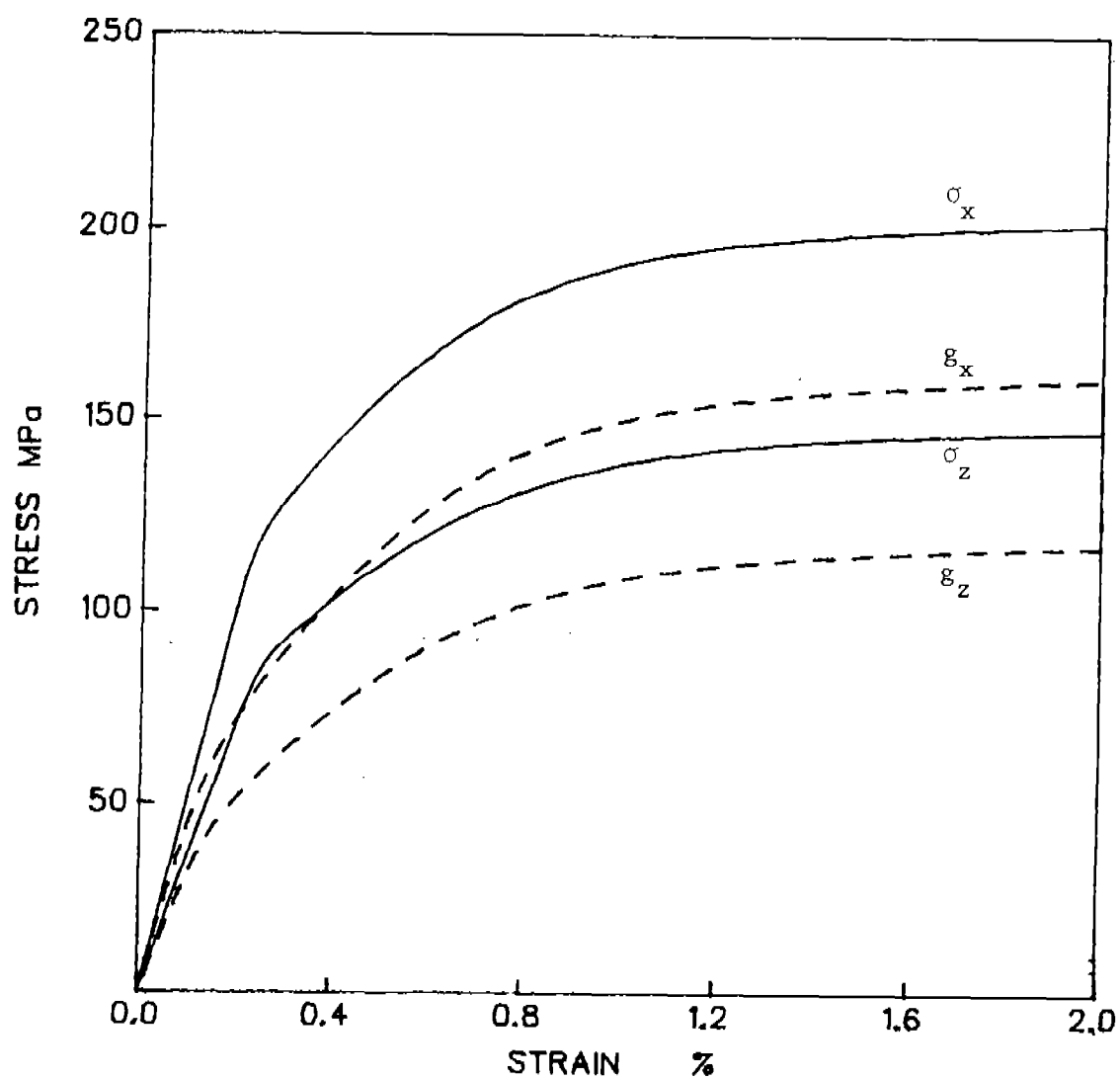


Figure 1. Uniaxial loading in x- and z-directions for $\dot{\epsilon} = 10^{-4}$
 (—) stress; (-----) equilibrium stress

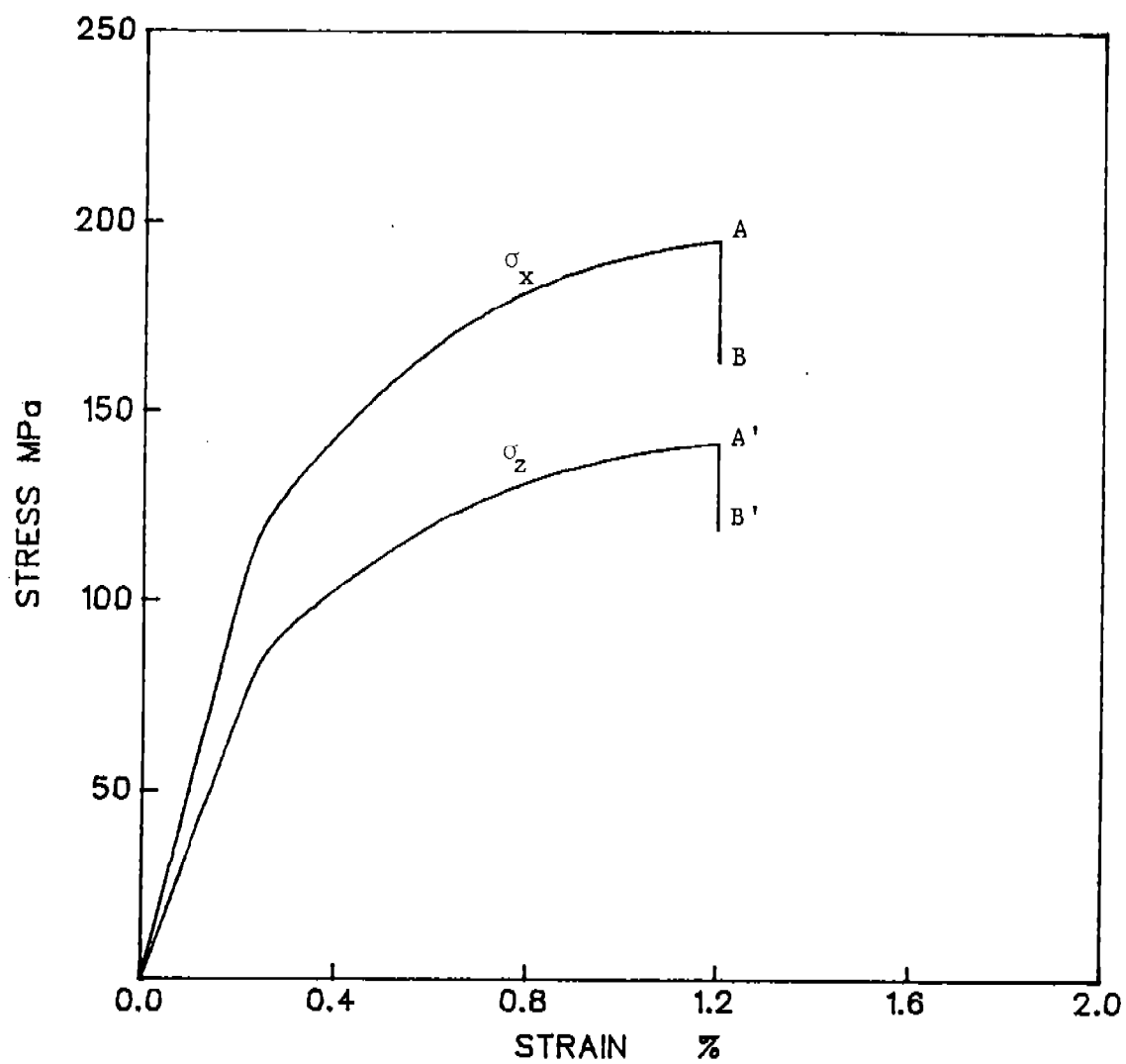


Figure 2. Uniaxial loading in x- and z-directions, at $\dot{\epsilon} = 10^{-4} \text{ s}^{-1}$, with 4 min. of relaxation periods at $\epsilon = 1.2\%$

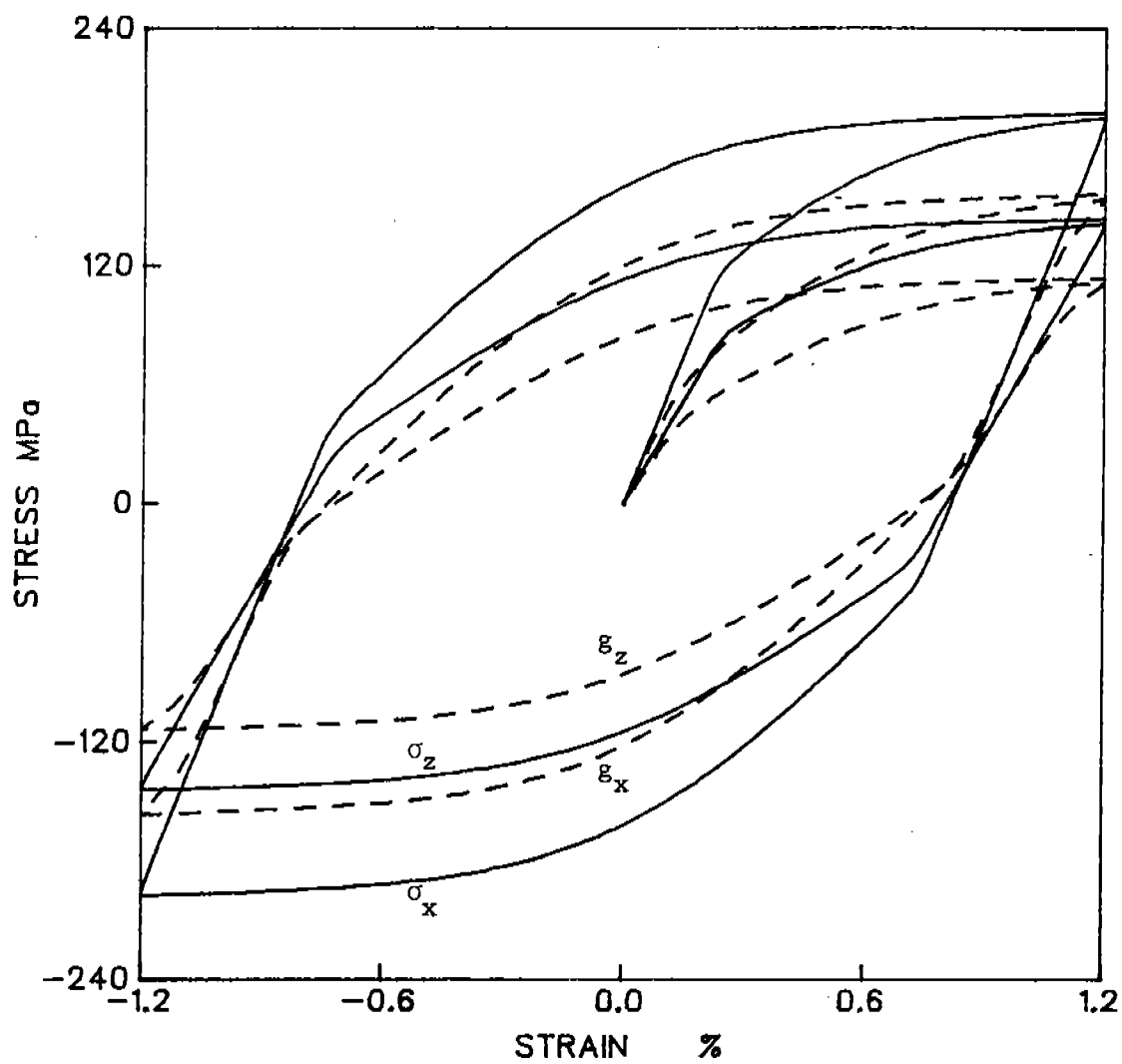


Figure 3. Cyclic uniaxial loading in x- and z-directions
 $\dot{\epsilon} = 10^{-4} \text{ s}^{-1}$

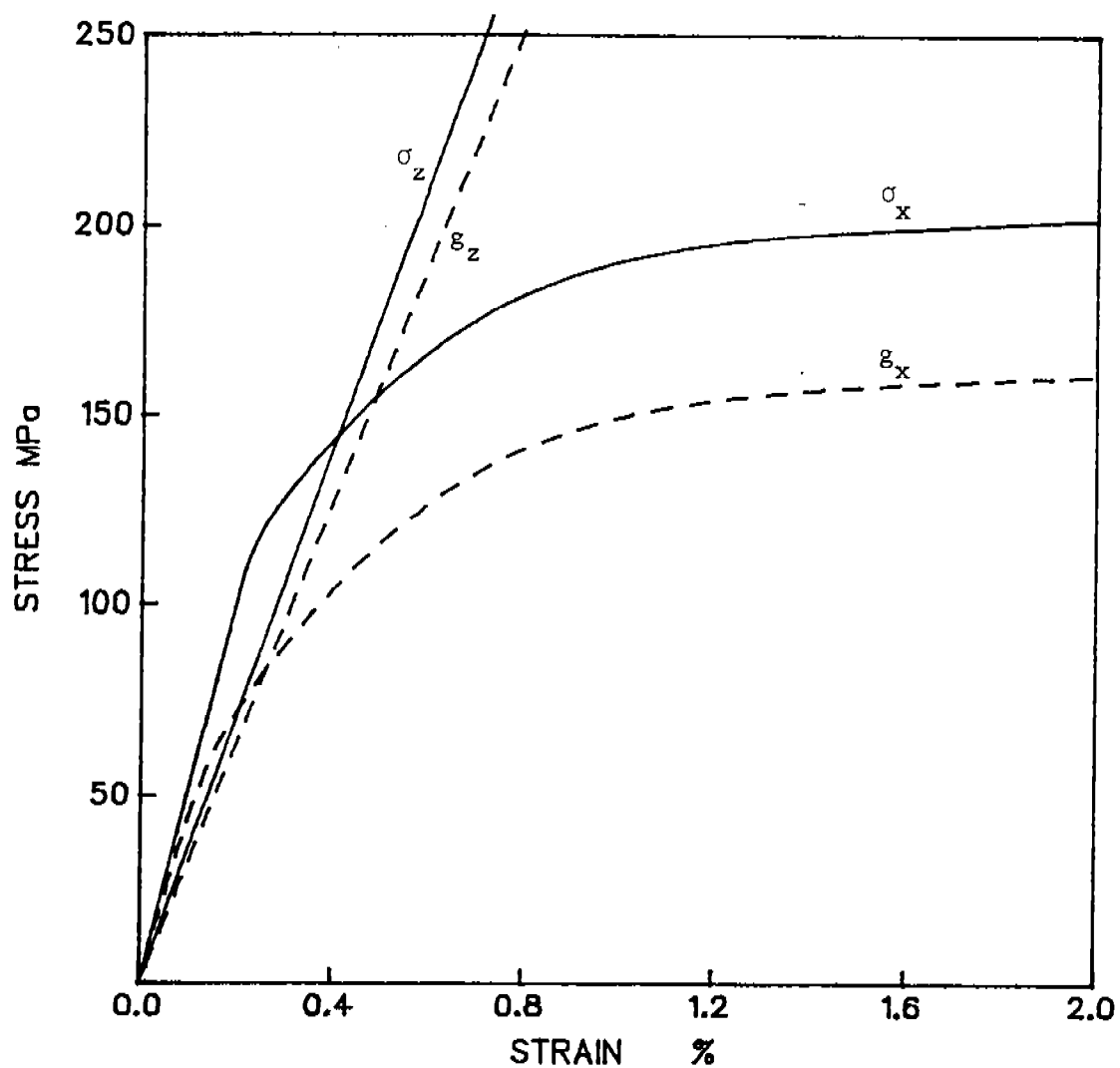


Figure 4. Uniaxial loading in x- and z-directions at $\dot{\epsilon} = 10^{-4} \text{ sec}^{-1}$.
 Nearly elastic behavior in the z-direction, $r_2 = 0$ in Γ and Γ_g .

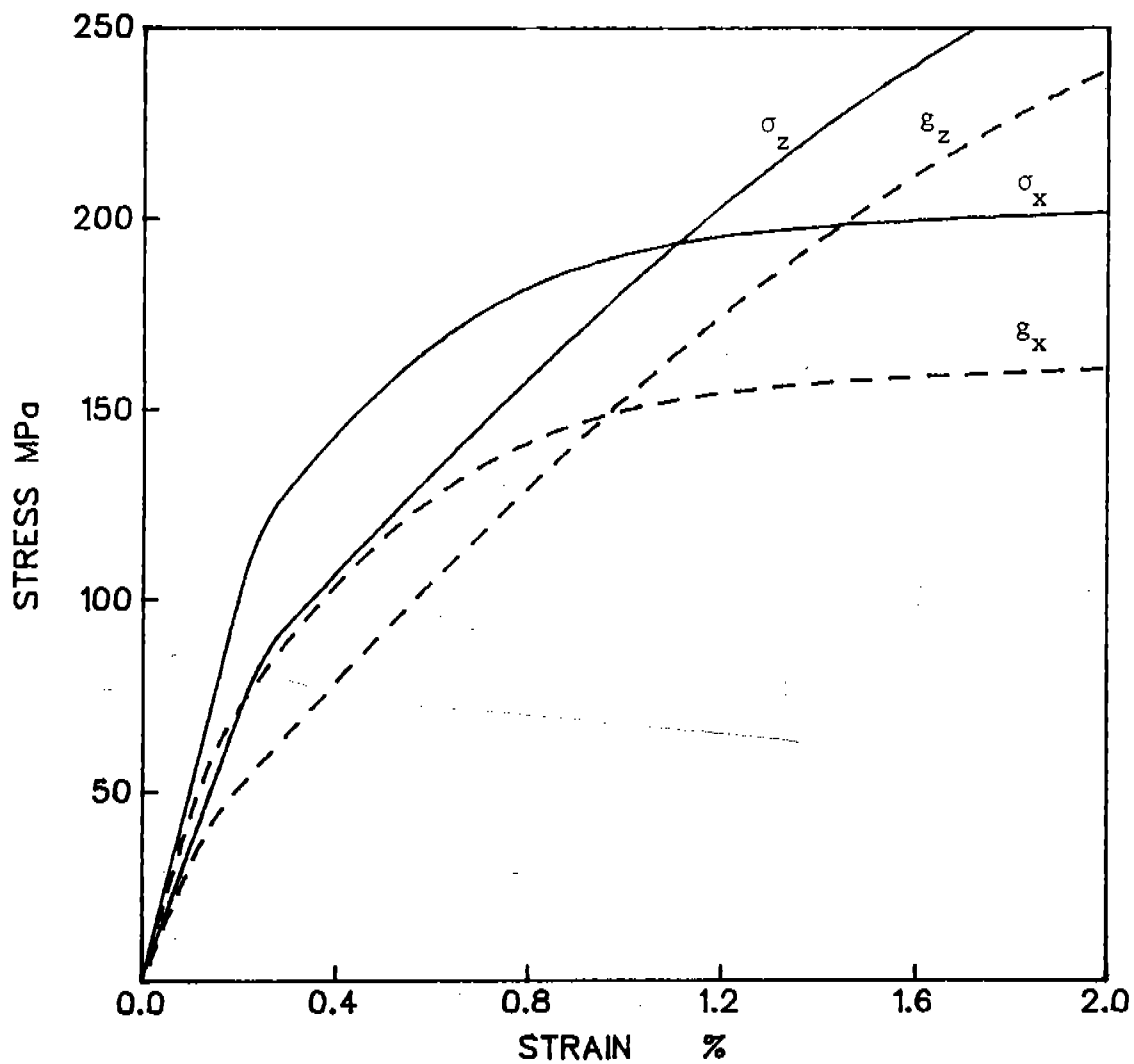


Figure 5. Uniaxial loading in the x- and z-directions at $\dot{\epsilon} = 10^{-4} \text{ sec}^{-1}$. Asymptotic level of the σ_z curve adjusted by choosing $r_2 = 0.5$ instead of E_1/E_3 , in Γ_g .

APPENDIX

The elastic coefficient tensor S_{ijpq} introduced in (5) can be defined implicitly by the following

$$\begin{aligned} \dot{\epsilon}_{ij}^e = & \bar{S}_1 m_{ij} + \bar{S}_2 n_{ij} + \bar{S}_3 l_{ij} + S_4 (m_{ip} \dot{\sigma}_{pj} + m_{jp} \dot{\sigma}_{pi}) \\ & + S_5 (n_{ip} \dot{\sigma}_{pj} + n_{jp} \dot{\sigma}_{pi}) + S_6 (l_{ip} \dot{\sigma}_{pj} + l_{jp} \dot{\sigma}_{pi}) \end{aligned} \quad (A.1)$$

where

$$\bar{S}_1 = \left((S_1 - 2S_4) m_{pq} + S_7 n_{pq} + S_8 l_{pq} \right) \dot{\sigma}_{pq} \quad (A.2a)$$

$$\bar{S}_2 = \left(S_7 m_{pq} + (S_2 - 2S_5) n_{pq} + S_9 l_{pq} \right) \dot{\sigma}_{pq} \quad (A.2b)$$

$$\bar{S}_3 = \left(S_8 m_{pq} + S_9 n_{pq} + (S_3 - 2S_6) l_{pq} \right) \dot{\sigma}_{pq} \quad (A.2c)$$

The representation given by (A.1) and (A.2) is taken from Boehler [18]. The elastic constants S_i , $i=1-9$ are related to the so-called engineering constants E_1, E_2, E_3 (Young's moduli), $\nu_{21}, \nu_{31}, \nu_{32}$ (Poisson's ratios), G_{23}, G_{12}, G_{13} (shear moduli). The relations are given as

$$S_1 = \frac{1}{E_1} \quad (A.3a)$$

$$S_2 = \frac{1}{E_2} \quad (A.3b)$$

$$S_3 = \frac{1}{E_3} \quad (A.3c)$$

$$S_4 = \frac{1}{4} \left(\frac{1}{G_{12}} + \frac{1}{G_{13}} - \frac{1}{G_{23}} \right) \quad (A.3d)$$

$$S_5 = \frac{1}{4} \left(\frac{1}{G_{23}} + \frac{1}{G_{12}} - \frac{1}{G_{13}} \right) \quad (A.3e)$$

$$S_6 = \frac{1}{4} \left(\frac{1}{G_{13}} + \frac{1}{G_{23}} - \frac{1}{G_{12}} \right) \quad (A.3f)$$

$$S_7 = -\frac{\nu_{21}}{E_2} \quad (\text{A.3g})$$

$$S_8 = -\frac{\nu_{31}}{E_3} \quad (\text{A.3h})$$

$$S_9 = -\frac{\nu_{32}}{E_3} \quad (\text{A.3i})$$

Equation (A.1) can also be put in matrix form

$$\begin{Bmatrix} \dot{\epsilon}_{11}^e \\ \dot{\epsilon}_{22}^e \\ \dot{\epsilon}_{33}^e \\ \dot{\epsilon}_{23}^e \\ \dot{\epsilon}_{13}^e \\ \dot{\epsilon}_{12}^e \end{Bmatrix} = \begin{bmatrix} S_{1111} & S_{1122} & S_{1133} & S_{1123} & S_{1113} & S_{1112} \\ & S_{2222} & S_{2233} & S_{2223} & S_{2213} & S_{2212} \\ & & S_{3333} & S_{3323} & S_{3313} & S_{3312} \\ & & & S_{2323} & S_{2313} & S_{2312} \\ \text{symmetric} & & & & S_{1313} & S_{1312} \\ & & & & & S_{1212} \end{bmatrix} \begin{Bmatrix} \dot{\sigma}_{11} \\ \dot{\sigma}_{22} \\ \dot{\sigma}_{33} \\ \dot{\sigma}_{23} \\ \dot{\sigma}_{13} \\ \dot{\sigma}_{12} \end{Bmatrix} \quad (\text{A.4})$$

where the S_{ijpq} are given by

$$\begin{aligned} S_{ijpq} = & (S_4 m_{ip} + S_5 n_{ip} + S_6 l_{ip}) \delta_{qj}^* \\ & + (S_4 m_{iq} + S_5 n_{iq} + S_6 l_{iq}) \delta_{pj}^* \\ & + (S_4 m_{jp} + S_5 n_{jp} + S_6 l_{jp}) \delta_{qi}^* \\ & + (S_4 m_{jq} + S_5 n_{jq} + S_6 l_{jq}) \delta_{pi}^* \\ & + m_{ij} \left((S_1 - 2S_4) m_{pq} + S_7 n_{pq} + S_8 l_{pq} \right) \\ & + n_{ij} \left(S_7 m_{pq} + (S_2 - 2S_5) n_{pq} + S_9 l_{pq} \right) \\ & + l_{ij} \left(S_8 m_{pq} + S_9 n_{pq} + (S_3 - 2S_6) l_{pq} \right) \end{aligned} \quad (\text{A.5})$$

The quantity δ_{ij}^* is related to the Kronecker δ_{ij} and assumes different values for different components of S_{ijkl} ,

$$\delta_{ij}^* = \frac{1}{2} \delta_{ij} \text{ for } \begin{cases} S_{1111}, S_{2222}, S_{3333}, \\ S_{1122}, S_{1133}, S_{2233} \end{cases} \quad (\text{A.5a})$$

and

$$\delta_{ij}^* = \delta_{ij} \text{ for } \begin{cases} S_{1113}, S_{1112}, S_{2223}, S_{2212} \\ S_{3323}, S_{3313}, S_{2323}, S_{2313} ; \\ S_{2312}, S_{1313}, S_{1312}, S_{1212} \end{cases} \quad (\text{A.5b})$$

for $S_{1123}, S_{2213}, S_{3312}$ either of the above definitions can be employed.

The expression on the right-hand side of (A.5) is not a tensorial representation*. It should be viewed as a formula for computing the components of the fourth-order tensor $\underline{\underline{S}}$ for a given orientation of the preferred directions M_i, N_i and L_i .

* A representation of the fourth-order tensor of orthotropy based on the geometry tensors $\underline{m}, \underline{n}$ and \underline{l} is given in [9]. It is extremely difficult to relate the nine constants of this representation to those of (A.5).

The inverse of (A.1), (or equation (7)) is

$$\begin{aligned} \dot{\sigma}_{ij} = & \bar{C}_1 m_{ij} + \bar{C}_2 n_{ij} + \bar{C}_3 l_{ij} + C_4 (m_{ip} \dot{\epsilon}_{pj}^e + m_{jp} \dot{\epsilon}_{pi}^e) \\ & + C_5 (n_{ip} \dot{\epsilon}_{pj}^e + n_{jp} \dot{\epsilon}_{pi}^e) + C_6 (l_{ip} \dot{\epsilon}_{pj}^e + l_{jp} \dot{\epsilon}_{pi}^e) \end{aligned} \quad (A.6)$$

where

$$\bar{C}_1 = \left((C_1 - 2C_4) m_{pq} + C_7 n_{pq} + C_8 l_{pq} \right) \dot{\epsilon}_{pq}^e \quad (A.7a)$$

$$\bar{C}_2 = \left(C_7 m_{pq} + (C_2 - 2C_5) n_{pq} + C_9 l_{pq} \right) \dot{\epsilon}_{pq}^e \quad (A.7b)$$

$$\bar{C}_3 = \left(C_8 m_{pq} + C_9 n_{pq} + (C_3 - 2C_6) l_{pq} \right) \dot{\epsilon}_{pq}^e \quad (A.7c)$$

The elastic constants C_i , $i=1,9$ are related to the engineering constants through

$$C_1 = \frac{E_1}{D} \left(1 - \nu_{32}^2 \frac{E_2}{E_3} \right) \quad (A.8a)$$

$$C_2 = \frac{E_2}{D} \left(1 - \nu_{31}^2 \frac{E_1}{E_3} \right) \quad (A.8b)$$

$$C_3 = \frac{E_3}{D} \left(1 - \nu_{21}^2 \frac{E_1}{E_2} \right) \quad (A.8c)$$

$$C_4 = G_{13} + G_{12} - G_{23} \quad (A.8d)$$

$$C_5 = G_{12} + G_{23} - G_{13} \quad (A.8e)$$

$$C_6 = G_{23} + G_{13} - G_{12} \quad (A.8f)$$

$$C_7 = \frac{E_2}{D} \left(\nu_{32} \nu_{31} \frac{E_1}{E_3} + \nu_{21} \frac{E_1}{E_2} \right) \quad (A.8g)$$

$$C_8 = \frac{E_3}{D} \left(\nu_{32} \nu_{21} \frac{E_1}{E_3} + \nu_{31} \frac{E_1}{E_3} \right) \quad (A.8h)$$

$$C_9 = \frac{E_3}{D} \left(\nu_{31} \nu_{21} \frac{E_1}{E_3} + \nu_{32} \frac{E_2}{E_3} \right) \quad (A.8i)$$

and

$$D = 1 - \nu_{21}^2 \frac{E_1}{E_2} - \nu_{31}^2 \frac{E_1}{E_3} - \nu_{32}^2 \frac{E_2}{E_3} - 2 \nu_{21} \nu_{31} \nu_{32} \frac{E_1}{E_3}, \quad (\text{A.9})$$

Equation (A.6) can also be written in the matrix form similar to (A.4). The components of C_{ijpq} can be obtained from (A.5) by simply replacing S_i by C_i given in (A.8), $i=1,9$.

It should be noted that there are some restrictions in the engineering constants E_1 , E_2 , ν_{21} , etc., and they are given by Jones [19].

The anisotropy ratios r_{ijpq} used in the text can be obtained from $(E_1 S_{ijpq})$ by replacing

$$\left. \begin{aligned} \frac{E_1}{E_2} &\text{ by } r_1 \\ \frac{E_1}{E_3} &\text{ by } r_2 \\ \frac{E_1}{2} \left(\frac{1}{G_{12}} + \frac{1}{G_{13}} - \frac{1}{G_{23}} \right) &\text{ by } r_3 \\ \frac{E_1}{2} \left(\frac{1}{G_{23}} + \frac{1}{G_{12}} - \frac{1}{G_{13}} \right) &\text{ by } r_4 \\ \frac{E_1}{2} \left(\frac{1}{G_{13}} + \frac{1}{G_{23}} - \frac{1}{G_{12}} \right) &\text{ by } r_5 \\ \nu_{21} &\text{ by } r_{21} \\ \nu_{31} &\text{ by } r_{31} \\ \nu_{32} &\text{ by } r_{32} \end{aligned} \right\} \quad (\text{A.10})$$

The inverse r_{ijpq}^{-1} is obtained from C/E_1 via the same procedure.

LINEAR STABILITY OF SHEAR FLOW OF A VISCOELASTIC FLUID

Yuriko Renardy and Michael Renardy
Mathematics Research Center
University of Wisconsin
610 Walnut St.
Madison, WI 53705

ABSTRACT. The stability of plane Couette flow of an upper convected Maxwell fluid is investigated using a numerical method. No evidence of instabilities has been found. The essential features of the linearized spectrum, as well as those of the numerical approximation, which can lead to artificial instabilities, are discussed.

INTRODUCTION. While there is a fairly comprehensive picture of the stability of parallel shear flows of Newtonian fluids, little is known about viscoelastic fluids. Experimentally, instabilities of flows of polymers are observed in a variety of geometries, even though the Reynolds number (a measure of the effect of inertia) is small. These instabilities are caused by the elasticity of the fluid, which is measured by another dimensionless parameter, called the Weissenberg or Deborah number. There seems to be considerable controversy over the precise origin of these instabilities, e.g. whether they are a feature of parallel shear flows or originate in inflow or outflow regions; for a review of experiments and various attempts at theoretical explanations see [2].

Denn and his coworkers [1], [3], [5], have investigated the linear stability of plane Poiseuille flow of an upper convected Maxwell fluid. In the Newtonian case, this flow becomes unstable at a critical Reynolds number $R_c \approx 5772$. Porteous and Denn [3] found that a small amount of elasticity decreases the critical Reynolds number to values of 1000-2000 when the Weissenberg number is increased to 1. At higher values of the Weissenberg number, they experienced numerical difficulties. These results prompted the conjecture that further increase of the Weissenberg number may continue to decrease the critical Reynolds number, and that there might ultimately be instabilities even at very low Reynolds numbers. While Rothenberger, McCoy, and Denn [5] claimed to have found such instabilities, a later study by Ho and Denn [1] indicated that these were an artifact of the discretization, and no evidence of real instabilities at low Reynolds number was found.

In this paper, we look at linear stability of plane Couette flow for an upper convected Maxwell fluid. In contrast to Denn and his coworkers, who used a shooting method to compute individual eigenvalues, we use a spectral method and a matrix eigenvalue solver to compute the whole spectrum of the problem. In this way we can obtain a more comprehensive picture not only of the spectrum of the continuous problem, but also of the behavior of the numerical approximation and of the origin and nature of artificial instabilities.

II. FORMULATION OF THE PROBLEM. We consider the flow between two parallel plates moving in opposite directions with equal velocities. The direction of the motion of the plates is the x-direction and the direction perpendicular to the plates in the y-direction. We non-dimensionalize the problem in such a way that the plates are at $y = 1$ and $y = -1$ and move in the x-direction with velocities $+1$ and -1 , respectively. The equations of the problem are the equation of motion

$$R \left[\frac{\partial \underline{u}}{\partial t} + (\underline{u} \cdot \nabla) \underline{u} \right] = -\nabla p + \text{div} \underline{T}$$

$$\text{div } \underline{u} = 0 ,$$
(1)

and the constitutive relation for an upper convected Maxwell fluid

$$\underline{T} + W \left[\frac{\partial \underline{T}}{\partial t} + (\underline{T} \cdot \nabla) \underline{T} - (\nabla \underline{u}) \underline{T} - \underline{T} (\nabla \underline{u})^T \right] = \nabla \underline{u} + (\nabla \underline{u})^T .$$
(2)

Here R is the Reynolds number and W the Weissenberg number. Plane Couette flow is the trivial solution

$$\underline{u} = (y, 0) , \quad \underline{T} = \begin{pmatrix} 2W & 1 \\ 1 & 0 \end{pmatrix} .$$
(3)

We linearize at this solution. Since Squire's theorem holds for the upper convected Maxwell fluid [6], only two-dimensional disturbances need to be considered. We separate variables and look at disturbances proportional to $e^{i\alpha x} e^{\sigma t}$, and we express the perturbed velocities in terms of a stream function. The resulting eigenvalue problem is [3]

$$\phi^{(4)} + b_3(y) \phi''' + b_2(y) \phi'' + b_1(y) \phi' + b_0(y) \phi = SR(yi\alpha + \sigma)(\phi'' - \alpha^2 \phi) ,$$
(4)

where

$$S = 1 + i\alpha W y + W \sigma ,$$
(5)

and

$$b_3(y) = 2 \frac{S'(S-1)}{S} ,$$

$$b_2(y) = -2\alpha^2 + 2 \frac{S'^2(S-1)^2}{S^2} ,$$

$$b_1(y) = -2\alpha \frac{S'(S-1)}{S} - 4 \frac{S'^3(S-1)}{S^2} ,$$
(6)

$$b_0(y) = \alpha^4 - 2\alpha^2 \frac{S'^2(S^2+1)}{S^2} + 4 \frac{S'^4}{S^2} ,$$

$$S' = \frac{dS}{dy} = i\alpha W .$$

The boundary conditions are

$$\phi = \phi' = 0 \text{ at } y = \pm 1 .$$

For the numerical treatment, we multiply (4) by S^2 and expand ϕ in a series of Chebyshev polynomials. This leads to a matrix eigenvalue problem of the form $\det (A - \sigma B) = 0$, which was solved by the NAG routine F02GJF. In our computations, we used between 24 and 60 Chebyshev polynomials.

III. Discussion of results. Before discussing our numerical results, we outline some features which can be obtained analytically. For $S = 0$, the equation has a singular point. This leads to a continuous spectrum given by

$$S = 0 \rightarrow \sigma = -\frac{1}{W} + i\alpha y, \quad y \in [-1, 1] . \quad (8)$$

If $S \neq 0$, the equation is regular, and eigenvalues are isolated. An exact solution to the problem can be obtained in the limiting case $\alpha = 0$. If we assume $S \neq 0$, we obtain $\phi^{(4)} = \sigma R(1 + \sigma W)\phi'' = : \sigma^* \phi''$. (9)

The eigenvalues σ^* are real and negative and tend to $-\infty$. We have

$$\sigma = \frac{-1 \pm \sqrt{1 + 4 \frac{W}{R} \sigma^*}}{2W} , \quad (10)$$

and we can conclude from this that, except for a finite number, the eigenvalues have real part $-\frac{1}{2W}$. Even for $\alpha \neq 0$, a formal asymptotic argument assuming σ large and very oscillatory eigenfunctions leads to the conclusion that the real parts of the eigenvalues should tend to $-\frac{1}{2W}$ asymptotically. The Newtonian eigenvalues are recovered from (10) if $\frac{W}{R} \sigma^*$ is small. That is, the spectrum looks more Newtonian if W is small and R is large.

The numerically computed eigenvalues reflect these overall features. No evidence of instability was found in the parameter range where we could obtain converged results. For W in excess of about 5, round-off errors caused substantial problems, which we hope to overcome in the future by using higher precision arithmetic. For the Newtonian case, it is known that plane Couette flow is always stable [4].

The presence of a continuous spectrum suggests the possibility of eigenvalues which either bifurcate from it or approach it in certain limits. Such eigenvalues were indeed found in our calculations. In the following table we have $R = 1$, $\alpha = 1$, and we show one of the eigenvalues as a function of W .

W	Re σ	Im σ
2.0	-0.370	± 0.774
1.7	-0.460	± 0.748
1.5	-0.543	± 0.730
1.2	-0.722	± 0.703
1.0	-0.900	± 0.687
0.5	-1.94	± 0.655
0.2	-4.97	± 0.646

Clearly, the real parts approach $-\frac{1}{W}$ as $W \rightarrow 0$.

It is interesting to look not only at converged results, but also at the behavior of the numerical approximations. It is well known that the scheme used here leads to "spurious" modes, i.e. to eigenvalues of the discrete problem which are pure artifacts of the discretization and do not approximate those of the differential equation even in a qualitative sense. In our calculations, we find four such spurious eigenvalues; they can be stable or unstable. The following discussion is concerned with the remaining eigenvalues, which approximate either the discrete or the continuous spectrum.

Since the scheme we use is infinite order accurate, a number of eigenvalues in the discrete spectrum are approximated very well provided we use a sufficient number of Chebyshev modes. As the imaginary parts of the eigenvalues increase, their real parts first approach $-\frac{1}{2W}$, reflecting the behavior of the continuous problem. However, for eigenvalues with large imaginary parts, the accuracy of the numerical approximation deteriorates, and the real parts of the computed eigenvalues move away from $-\frac{1}{2W}$. If W is sufficiently large, they may become positive, thus leading to artificial instabilities. It is not clear that using more Chebyshev polynomials will remove those artificial instabilities: For any fixed range of $\text{Im}\sigma$, the approximation is improved, but at the same time, eigenvalues with higher and higher imaginary parts are introduced, and the instability may just be shifted to those higher modes.

The continuous spectrum extends from $-\frac{1}{W} - i\alpha$ to $-\frac{1}{W} + i\alpha$. Since the infinite order accuracy of the method applies only to isolated eigenvalues with C^∞ eigenfunctions and not to the continuous spectrum, the approximation is relatively poor. It is better near the ends and worse towards the middle. For high values of W and/or α , this leads again to artificial instabilities. Since these artificial instabilities occur near $\text{Im}\sigma = 0$, they are particularly damaging if one tries to use this kind of scheme for transient calculations: Unstable eigenvalues with large imaginary parts, such

as those resulting from the numerical approximation of the discrete spectrum, can be stabilized by using appropriate time discretizations, but this is not possible for eigenvalues close to zero. We suspect that the artificial instabilities found by Denn and his coworkers [1], [5] also result from poor approximations to the continuous spectrum.

As expected, the quality of the numerical approximation deteriorates with increasing W . Also, at the same time, the potential for artificial instabilities is increased due to fact that the real parts of the eigenvalues are proportional to $-\frac{1}{W}$, i.e. there is less to "spare" if you want to remain stable. Higher Reynolds numbers tend to improve the features of the numerics.

Calculations of steady flows of viscoelastic fluids have so far had remarkably little success. There have been suggestions that flows could be computed by integrating the equations forward in time and waiting for the calculations to settle down to a steady state. The features found in this study suggest that great caution is advisable with this approach. We found artificial instabilities resulting not only from spurious modes, but also from poorly approximated ones, and similar problems were encountered by Denn with a different numerical scheme. It remains an open question as to whether schemes can be constructed to avoid such problems.

References

- [1] T.C. Ho and M.M. Denn, Stability of plane Poiseuille flow of a highly elastic liquid, J. Non-Newtonian Fluid Mech. 3(1977/78), pp. 179-195.
- [2] C.J.S. Petrie and M.M. Denn, Instabilities in polymer processing, AIChE J. 22(1976), pp. 209-236.
- [3] K.C. Porteous and M.M. Denn, Linear Stability of plane Poiseuille flow of viscoelastic liquids, Trans. Soc. Rheology 16(1972), pp. 295-308.
- [4] V.A. Romanov, Stability of plane-parallel Couette flow, Functional Anal. and Its Appl. 7(1973), pp. 137-146.
- [5] R. Rothenberger, D.H. McCoy, and M.M. Denn, Flow instability in polymer melt extrusion, Trans. Soc. Rheology 17(1973), pp. 259-269.
- [6] G. Tlapa and B. Bernstein, Stability of a relaxation-type viscoelastic fluid with slight elasticity, Phys. Fluids 13(1970), pp. 565-568.

EFFECT OF A WALL ON THE LIFT FORCE

Donald A. Drew
Department of Mathematical Science
Rensselaer Polytechnic Institute
Troy, New York U.S.A. 12180-3590

Introduction

Particle-fluid flows occur in filtration processes and wear and erosion. Particle motion near a wall is a complex question. Whether a particle actually touches a wall is complicated by short range forces and inhomogeneities in geometry and materials. A particle must first move through the fluid before short range forces dominate. For this reason, we study the forces on a particle moving in a shearing fluid near a wall.

In this paper, we derive the forces on a spherical particle in Poiseuille flow in a channel, accounting for the inertial force on the particle due to the presence of the wall. We use the method of Fourier transform, first used by Childress (1964) for this type of inertial force calculation. It was applied to shear flows by Saffman (1964). The method was generalized to general shapes by Harper and Chang (1968), and to straining and rotating flows by Drew (1978).

Equation of Motion

The equations of motion for a sphere moving near a wall in an incompressible viscous fluid of density ρ and viscosity μ are

$$\nabla \cdot \underline{v} = 0 \quad (1)$$

$$\rho \left[\frac{\partial \underline{v}}{\partial t} + \underline{v} \cdot \nabla \underline{v} \right] = -\nabla p + \mu \nabla^2 \underline{v} \quad (2)$$

in $\Omega(t)$, which is defined to be

$$\Omega(t) = \left\{ \underline{r} \mid r_3 \geq 0 \text{ and } |\underline{r} - \underline{r}_p(t)| \geq a \right\} . \quad (3)$$

Here \underline{v} is the velocity of the fluid, p is the pressure, $\underline{r} = (r_1, r_2, r_3)$ is the spatial position, t is time, \underline{r}_p is the center of the sphere, and a is the radius of the sphere. We assume $r_{p3}(t) = d(t) \geq a$, so that the sphere does not intersect the wall.

This work was supported by U.S. Army Research Office Contract DAAG29-82-K-0185

The boundary conditions are

$$\underline{v} = 0 \quad (4)$$

at $r_3 = 0$, and

$$\underline{v} = \frac{dr_p}{dt} + \underline{\Omega}(t) \times (\underline{r} - \underline{r}_p) \quad (5)$$

on $|\underline{r} - \underline{r}_p(t)| = a$. Here $\underline{\Omega}(t)$ is the angular velocity of the sphere. The equations of motion for the sphere are

$$m \frac{d^2 \underline{r}_p}{dt^2} = \underline{F} + \iint_{S_p} \underline{n} \cdot \left[-p \underline{I} + \mu (\nabla \underline{v} + (\nabla \underline{v})^T) \right] ds \quad (6)$$

$$I \frac{d\underline{\Omega}}{dt} = \underline{I} + \iint_{S_p} (\underline{r} - \underline{r}_p) \times \underline{n} \cdot \left[\mu (\nabla \underline{v} + (\nabla \underline{v})^T) \right] ds \quad (7)$$

where m is the mass of the sphere, I is its moment of inertia, \underline{F} is the externally applied force, and \underline{I} is the externally applied torque.

Assume the sphere is small, and the flow far from the sphere is quadratic. Specifically, the flow in the absence of the sphere is given by

$$\underline{v}_0 = \underline{e}_1 \frac{1}{\mu} \frac{dp}{dr_1} \left[\frac{(r_1)^2}{2} - \frac{1}{2} r_1^2 \right], \quad p_0 = P_0 + P_1 r_1 \quad (8)$$

Here $\frac{1}{2} \gg a$, and $P_1 \neq 0$. With no loss of generality we set $P_0 = 0$. The velocity of the fluid relative to the sphere is

$$\underline{u} = \underline{v}_0 - \frac{dr_p}{dt} \quad (9)$$

We shall call eq. (8) the outer solution. This flow is steady. Note that this solution does not satisfy the boundary conditions at the particle surface.

Inner Flow

Let us consider the inner problem. The velocity is scaled by $U = |\underline{v}_0(\underline{r}_p) - d\underline{r}_p/dt|$, and the length scale is a . We define

$\kappa = a/d(0)$, and $Re = \rho U a / \mu$. We shall refer to Re as the Reynolds number, and assume that κ and $Re \ll 1$.

Note that the flow is unsteady due to the relative motion of the wall and the changing motion of the fluid far from the sphere, as viewed from the sphere. The motion of the coordinate system is assumed slow, so that we can neglect the non-inertial accelerations (Coriolis force, centripetal force).

The imposition of the wall boundary condition depends on

the size of κ . If $\kappa = O(1)$, then the wall influences the flow (and hence the motion of the sphere) strongly. For smaller κ , we must ascertain whether wall effects or other perturbation effects are more important.

For small κ , the solution the inner problem is given by Brenner (1964). The details are omitted, but the salient features will be summarized here. The force and torque on the sphere are given by Faxen's laws. The force is

$$\underline{F}_i = 6\pi\mu a (\underline{v}_0(\underline{r}_p) - d\underline{r}_p/dt) - \frac{2}{3}\pi a^3 \nabla p_0 \quad (10)$$

If we assume no net torque on the sphere, we have

$$\underline{\Omega} = \frac{1}{2} \nabla \times \underline{v}_0 \quad (11)$$

The inner flow is equal to the outer flow, plus a Stokeslet, plus terms which decay faster than $|\underline{r}_i|^{-2}$ as $|\underline{r}_i| \rightarrow \infty$.

Inertial Region

Let us consider the flow on a scale between the inner and outer. Harper and Chang (1968) give a careful explanation of the balances which must occur. In this region the appropriate Reynolds number is one based on the velocity gradient,

$$Re_S = \frac{l \nabla v_0 a^2}{\mu}$$

Roughly, on any length scale between a and $a Re_S^{-1/2}$, the flow consists of the outer flow, plus a Stokeslet, to lowest order.

For length scales between $a Re_S^{-1/2}$ and l , the flow is the outer flow, to lowest order. A critical balance occurs on a length scale equal to $a Re_S^{-1/2}$. At this scale, inertia contributes a perturbation in the flow of order $Re_S^{1/2}$, which by virtue of Faxen's laws, give an $O(Re_S^{1/2})$ contribution to the force.

The approximate equations valid in the inertial subregion are

$$\nabla \cdot \underline{v} = 0 \quad (12)$$

$$\rho U_S \left[r_3 \frac{\partial \underline{v}}{\partial r_1} + v_3 \underline{e}_1 \right] = - \nabla p + \mu \nabla^2 \underline{v} \quad (13)$$

$$\underline{v} = 0 \text{ on } r_3 = -d \quad (14)$$

Where $U_S = \frac{1}{\mu} \frac{dp_0}{dr_1} (d - l)$, and for convenience, we have translated the origin to the sphere center. The solution must match to the Stokeslet from the inner solution.

We can also see the relation between the distance from

boundary and the inertial force. If $d \text{Re}_S^{1/2} \ll a \ll d$, the inner flow "sees" the boundary. The inner problem may be solved by perturbation methods (Cox and Hsu 1977). If $a \ll d \text{Re}_S^{1/2}$, the inertial correction due to this inertial flow will occur at $O(\text{Re}_S)^{1/2}$ and will dominate the $O(\kappa)$ corrections. In this case, the analysis of Saffman (1965, 1968), generalized by Harper and Chang (1968) gives the $O(\text{Re}_S)^{1/2}$ correction to the force. When $a = O(d \text{Re}_S^{1/2})$, the force can be computed from (12-14). To compute the force on the sphere, we must compute the induced velocity at the origin. Then

$$\underline{F} = 6\pi\mu a \underline{v}^1(0). \quad (15)$$

Harper and Chang (1968), following Saffman (1965; 1968), found the solution to (12-13) in the infinite domain. Their solution also matches to the Stokeslet at $\underline{r} = 0$. If we denote the solution in the infinite domain as \underline{v}_0, p_0 , and write

$$\underline{v} = \underline{v}_0 + \underline{v}_1, \quad (16a)$$

$$p = p_0 + p_1, \quad (16b)$$

$$\underline{r}' = \underline{r} / a \text{Re}_S^{1/2}, \quad (16c)$$

$$d' = d / a \text{Re}_S^{1/2}, \quad (16d)$$

we have

$$\nabla \cdot \underline{v}_1 = 0, \quad (17)$$

$$\left[r_3 \frac{\partial \underline{v}_1}{\partial r_1} + v_{13} \underline{e}_1 \right] = -\nabla p_1 + \mu \nabla^2 \underline{v}_1, \quad (18)$$

$$\underline{v}_1 = -\underline{v}_0 \text{ on } r_3 = -d, \quad (19)$$

where we drop the prime on \underline{r} and d .

If we apply a Fourier transform in r_1 and r_2 , we have

$$v_1 = \frac{1}{4\pi^2} \iint \underline{\Gamma}_1(\underline{k}, r_3) e^{i\hat{\underline{k}} \cdot \hat{\underline{r}}} d\hat{\underline{k}}, \quad (20a)$$

$$p_1 = \frac{1}{4\pi^2} \iint \pi_1(\underline{k}, r_3) e^{i\hat{\underline{k}} \cdot \hat{\underline{r}}} d\hat{\underline{k}}, \quad (20b)$$

where $\hat{\underline{r}} = (r_1, r_2, 0)$, $\hat{\underline{k}} = (k_1, k_2, 0)$. If Fourier transforms are taken of (17,18), and the pressure transform π_1 is eliminated, a differential equation can be derived for the \underline{e}_3 component of $\underline{\Gamma}_1$. It is this velocity which gives rise to transverse, or lift forces. The equation is

$$\frac{\partial^4 \Gamma_{13}}{\partial r_3^4} - (2k^2 + i k_1 r_3) \frac{\partial^2 \Gamma_{13}}{\partial r_3^2} + (\hat{k}^4 + i k_1 r_3 \hat{k}^2) \Gamma_{13} = 0 \quad (21)$$

If we let $G_{13} = \partial^2 \Gamma_{13} / \partial r_3^2 - \hat{k}^2 \Gamma_{13}$, we have

$$\frac{\partial G_{13}}{\partial r_3^2} - (\hat{k}^2 + ik_1 r_3) G_{13} = 0 \quad (22)$$

Eq. (22) is an Airy equation. The solution for Γ_{13} is

$$\Gamma_{13} = ae^{-\hat{k}r_3} + bf(r_3) \quad (23)$$

where

$$f(r_3) = \int_0^{r_3} 2 \cosh [\hat{k}(r_3 - s)] G_{13}(s) ds \quad (24)$$

where $G_{13}(s)$ is the solution of (22) which decays as $s \rightarrow \infty$. This solution is found by Generalized Laplace transforms. The initial conditions for Γ_{13} are determined from eq. (16a) in conjunction with Harper and Chang's solution. We have

$$\Gamma_{13}(\hat{k}, -d) = - \int_{-\infty}^{\infty} \Gamma_{031} e^{-ik_3 d} dk_3 \quad (25)$$

$$\begin{aligned} \frac{\partial \Gamma_{13}}{\partial r_{13}}(\hat{k}, -d) &= -ik \cdot \Gamma_1(\hat{k}, -d) \\ &= i \int_{-\infty}^{\infty} (k_1 \Gamma_{011} + k_2 \Gamma_{021}) e^{-k_3 d} dk_3 \end{aligned} \quad (26)$$

where Γ_{011} , Γ_{021} , Γ_{031} are the components of the Fourier transform of \underline{v}_0 . They are given by Harper and Chang (1968).

The force in the e_3 direction calculated by Saffman and Harper and Chang is

$$F_D = 6.46 \quad (27)$$

The $O(Re_S)^{1/2}$ correction to the force in the e_3 direction equal to

$$F_1 = 6\pi\mu av_{13}(0) = \iint \Gamma_{13}(\underline{k}, 0) d\underline{k} \quad (28)$$

The integrals involved in finding Γ_{13} and subsequently F_1 have been approximated numerically using NAG library routine with truncation. The calculations were expensive and required large amounts of computer time and memory. We speculate whether a more direct approach might have been better. The result for $F = F_1\mu av_{13}(0)$ is shown in Figure 1 as a function of d . The total lift force is obtained by adding (27) and appropriate value from Figure 1, and multiplying by $(Re_S)^{1/2}$.

Conclusions

The lift force due to the slip-shear interaction is enhanced by the presence of a wall. The contribution due to the wall, shown in Figure 1, decreases with distance from the wall. The wall correction is quite small compared with the the slip-shear interaction force at moderate distances from the wall ($d = O(a Re^{-1/2})$).

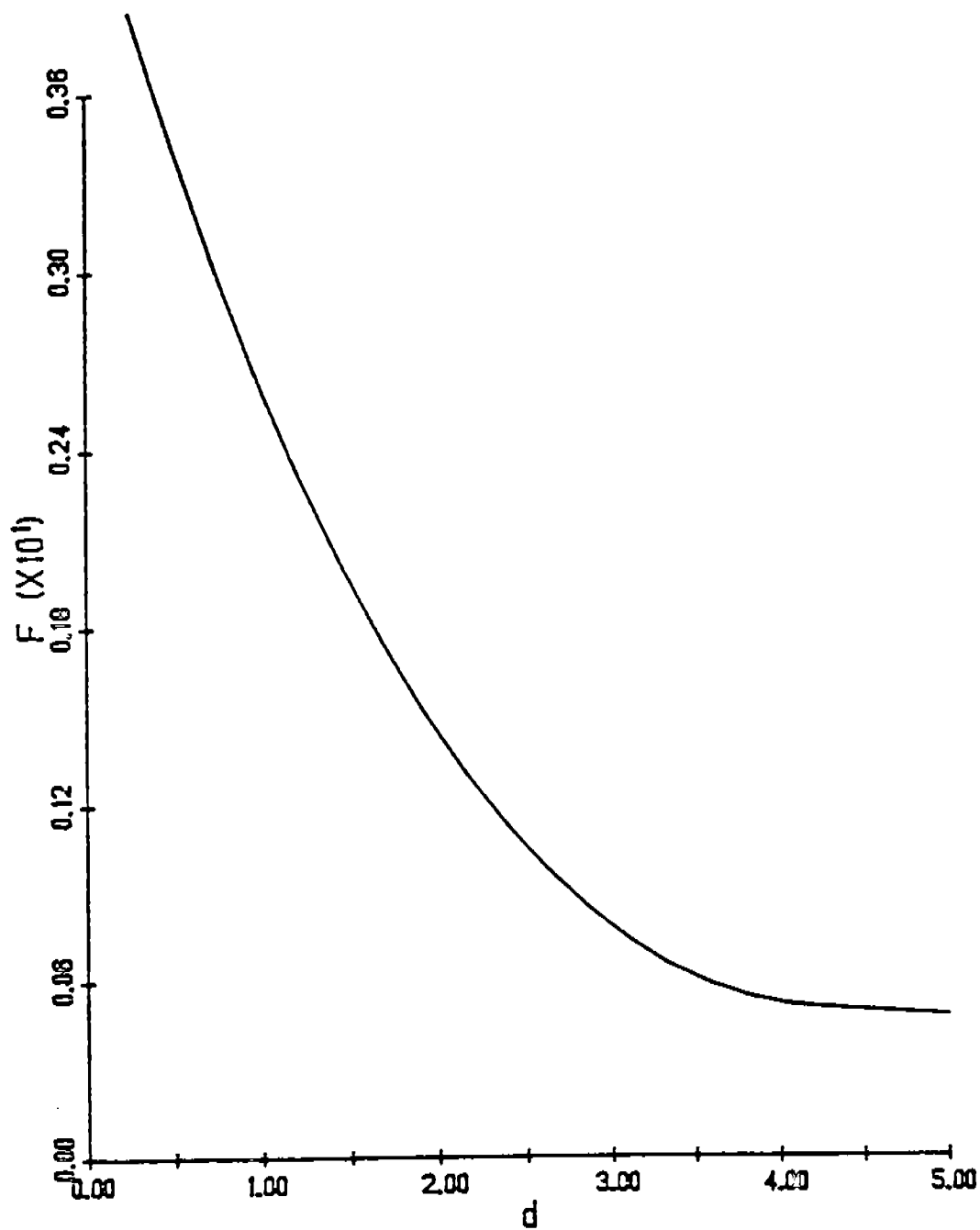


Figure 1. Dimensionless wall-induced force versus dimensionless distance from wall.

The above analysis presumes a force F on the sphere in axial direction. If no force acts on the sphere, the analysis gives a force of zero to the order computed, since the Stokeslet contribution in the inner flow is zero. With $F_i = 0$ and $T_i = 0$, the next contribution to the force is the strainlet (Schonberg, Drew and Belfort 1985). This alters the matching condition as $r \rightarrow 0$ for eqs. (17) and (18), and leads to $o(Re^{3/2})$ corrections (Lin, Peery and Schowalter 1970).

References

- Brenner, H. 1964 Chem. Eng. Sci. 19, 703.
- Childress, W.S. 1964 J. Fluid Mech. 20, 305.
- Cox, R.G. and Brenner, H. 1968 Chem. Eng. Sci. 23, 147.
- Cox, R.G. and Hsu, S.K. 1977 Int. J. Multiphase Flow 3, 201.
- Drew, D.A. 1978 J. Fluid Mech. 88, 393.
- Happel, J. and Brenner, H. 1965 Low Reynolds Number Hydrodynamics Prentice Hall.
- Harper, E.Y. and Chang, I.D. 1968 J. Fluid Mech. 33, 209.
- Lin, C-J., Peery, J.H., and Schowalter, W.R., 1970 J Fluid Mech. 44, 1.
- Saffman, P.G. 1965 J. Fluid Mech. 22, 385.
- Saffman, P.G. 1968 J. Fluid Mech. 31, 624.
- Schonberg, J.A., Drew, D.A. and Belfort, G. 1985 submitted for publication.
- Stokes, G.G. 1851 Trans. Camb. Phil. Soc. 9, 8.

THE EFFECTS OF NON-SPHERICITY AND RADIATIVE ENERGY LOSS ON THE
MIGRATION OF THE GAS BUBBLE FROM UNDERWATER EXPLOSIONS

K.C. Heaton
Weapon Systems Section, Armaments Division
Defence Research Establishment Valcartier
P.O. Box 8800, Courcellette
Quebec, GOA 1R0

ABSTRACT

One of the more important phenomena associated with the upward motion of a gas bubble from an underwater explosion is the significant departure from sphericity near the times of the minimum bubble radius. Neglecting this change in shape results in the prediction of a much faster upward velocity than actually occurs. The inclusion of this effect in the equations of motion has been exceedingly difficult because of the large magnitude of the departure from sphericity.

In this work, the shape of the bubble is described by an ellipsoid whose axes are allowed to vary independently, thus modelling, to first order, the changes of bubble shape. The Lagrangian equations of motion, incorporating the effects of the change of shape and of energy loss by the radiation of sound, are derived and solved for the case of a spheroidal bubble. The results of these calculations for various initial conditions are compared with analogous cases for a spherical bubble.

It is found that the spheroidal bubble model predicts a reduction in the upward translational motion of the bubble of a factor of approximately 2. A comparison of the predicted upward motion of a spheroidal bubble produced by 227.27 kg. of TNT detonated 46 metres below the surface shows good agreement with that which has actually observed.

I. INTRODUCTION

The formation of a bubble of gaseous detonation products always accompanies an underwater explosion. This bubble rises toward the surface of the water, responding as it does so to the change in the external pressure distribution with oscillatory motion, during the course of which it loses some of its energy through the emission of sonic pulses. Although the bubble is initially spherical, the effect of its upward motion is to distort it into a non-spherical shape, which becomes most pronounced in the neighbourhood of the minimum radius. The alteration in the bubble's shape further affects both the pulsations and the upward translational motion of the bubble. By means of finite element techniques, the equations of motion of the bubble can be solved, taking into account the effects of the changing shape of the bubble, although the amount of computing time required by this method limits its utility. Finite element methods have a further disadvantage in that physical insights into the systems considered are rather more difficult to come by than might otherwise have been the case.

Herring (1942) and others (eg. Taylor (1942) and Shiffman and Friedman (1944)) have treated the problem of the motion of the bubble by considering it to be a perfect sphere throughout its entire motion. This treatment yields values for the periods of radial pulsations of the bubble which are in good agreement with experimental data, but predicts a much more rapid movement toward the surface than is actually observed. This arises because the largest upward velocities of the bubble occur at those times when the bubble is near its minimum radius; it is precisely then that the largest departures from sphericity occur. Penney and Price (1942) and Ward (1943) included the effects of the non-sphericity of the bubble on its motion. However, it was always explicitly assumed in their derivations of the velocity potential of the flow about the bubble that the departures from sphericity are always small. Accordingly, their equations are not applicable near the times at which the bubble is at its smallest volumes.

Hicks (1972) was able to bring the value for the upward translational velocity of a spherical bubble at a minimum radius into agreement with experimental data by adding a drag term to the equations of motion. In his formulation, the drag coefficient is an empirical correction whose value is chosen to make the predicted rate of rise at the first minimum radius consistent with observation. However, a different drag coefficient must be selected for each charge mass and depth, requiring a comprehensive data base from which the appropriate value can be chosen for each case. For these reasons, a model for the bubble in which large deviations from sphericity and their effects on its translational motion are treated would be of considerable practical and theoretical interest.

In this work, a Lagrangian is derived for a bubble whose shape is not constrained to be always spherical but may become ellipsoidal as it moves upwards. The equations of motion for a general ellipsoidal bubble, incorporating the effects of loss of energy by radiation, are presented. The algorithm by which these equations of motion were solved numerically is briefly discussed. Computational results for some charge masses and depths are presented and compared with experimental data.

II. REVIEW OF PREVIOUS WORK

Taylor (1942) derived equations describing the motion of a spherical bubble of gas undergoing both radial pulsations and translational motion toward the water surface. These are:

$$2\pi\rho a^3 \left(\frac{da}{dt}\right)^2 + \frac{\pi}{3}\rho a^3 U^2 + \frac{4\pi}{3}\rho g z a^3 \quad [2.1]$$

$$= Y_0 - E(a)$$

$$U = -\frac{dz}{dt} = \frac{2g}{a^3} \int_0^t a^3 dt \quad [2.2]$$

where a is the radius of the bubble as a function of the time t , U its upward velocity, z the position of the bubble below the pressure datum (i.e. below the zero pressure level), $E(a)$ the internal energy of the gas comprising the bubble, g the gravitational acceleration, Y_0 the total energy of the bubble and ρ the density of the water. In Taylor's formulation, there was no mechanism included for energy loss, and hence Y_0 was taken to be a constant. For TNT explosions, it has been found (Herring 1942) that approximately 50% of the total explosion energy is retained by the bubble; in that case,

$$Y_0 = (1.85 \times 10^{10})M \quad [2.3]$$

where Y_0 is measured in ergs, and M , the original mass of the explosive charge, is given in gm. If one assumes that the gaseous explosion products obey the ideal gas law, then the internal pressure, P , is given by

$$P = k(\rho_g)^\gamma \quad [2.4]$$

where ρ_g is the density of the explosion products and γ the ratio of specific heats. Assuming that the entire mass, M , of the explosive has been converted to gas,

$$\rho_g = \frac{M}{\left(\frac{4\pi}{3}\right)a^3} \quad [2.5]$$

and hence,

$$E(a) = \int_a^{\infty} P dV \quad [2.6]$$

$$= \frac{kM\gamma a^{-3(\gamma-1)}}{(\gamma-1) \left(\frac{4\pi}{3}\right)^{\gamma-1}}$$

where $dV = 4\pi a^2 da$. Taylor, using the work of Jones, set

$$k = 7.83 \times 10^9 \quad [2.7]$$

$$\gamma = 1.25$$

for TNT where, in eq. [2.6], $E(a)$ is measured in ergs and m in gm.

Hicks (1972) incorporated a drag force, F_D , into the equations of motion, where F_D is given by

$$F_D = \frac{1}{2} C_D \rho a^2 U^2 \quad [2.8]$$

and the value of the drag coefficient, C_D , was chosen to be $C_D = 2.25$ in order to bring the distance travelled upward by the bubble at its first maximum into agreement with that actually observed for 500 lbs. of TNT detonated 150 ft. below the surface. By differentiating eq. [2.2], he obtained the rate of change of momentum with respect to time. The incorporation of F_D into the equation of motion yielded:

$$\frac{d}{dt}(a^3 U) = 2a^3 g - \frac{3}{4} C_D a^2 U^2 \quad [2.9]$$

L , the Lagrangian of the bubble, is given by

$$L = 2\pi\rho a^3 \left(\frac{da}{dt}\right)^2 + \frac{\pi}{3} \rho a^3 U^2 - \frac{4\pi}{3} a^3 \rho g z - E(a) \quad [2.10]$$

Hence, a more general form for the equations of motion of a spherical bubble is given by

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{a}} \right) - \frac{\partial L}{\partial a} = Q_a,$$

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{z}} \right) - \frac{\partial L}{\partial z} = Q_z \quad [2.11]$$

where $\dot{a} = \frac{da}{dt}$, $\dot{z} = \frac{dz}{dt} = -U$, and the Q_i are the generalised dissipative forces.

The dissipative drag force, Q_z , is given by F_D in eq. [2.8]. It has previously been shown (Heaton 1984) that Q_a , the generalised force for dissipation by radiation of sound, is given by

$$Q_a = -16\pi\rho a^2 \frac{\dot{a}^3}{C_S} - \frac{1}{C_S} \Delta Q_a \quad [2.12]$$

where

$$\Delta Q_a = 4\pi\rho a^2 [4a\ddot{a}\ddot{a} + (a^2\ddot{a}^2/\dot{a})] \quad [2.13]$$

$\ddot{a} = \frac{d^2 a}{dt^2}$, and C_S is the speed of sound in the water.

The terms contained in ΔQ_a are analogous to the radiation reaction terms in electromagnetic theory, and hence can be ignored in a first approximation, although this approach will underestimate the energy loss near the minimum radius, and overestimate it elsewhere.

Ward (1943) and Penney and Price (1942) derived equations of motion for a nearly spherical bubble by expanding the velocity potential, Φ , of the flow about the bubble in terms of the Legendre polynomials, P_n , thusly:

$$\Phi = \frac{A}{r} + B_1 \frac{P_1(\cos\theta)}{r^2} + B_2 \frac{P_2(\cos\theta)}{r^3} + \dots \quad [2.14]$$

where the coefficients A , B_1 and B_2 are functions only of time. They further assumed that the radius vector, $R(t)$, from the centre of the bubble to a point on its surface could be written as

$$R(t) = a + b_2 P_2(\cos\theta) + b_3 P_3(\cos\theta) + \dots \quad [2.15]$$

where a , b_2 , and b_3 are functions of time only. At the surface of the bubble,

$$\begin{aligned} \frac{dR}{dt} &= - \left(\frac{\partial \Phi}{\partial r} \right)_R - U \cos\theta \\ &= \frac{A}{R^2} + 2 \frac{B_1}{R^3} P_1(\cos\theta) + 3 \frac{B_2}{R^4} P_2(\cos\theta) \end{aligned} \quad [2.16]$$

Substituting eq. [2.15] into eq. [2.16] and equating the coefficients of the Legendre polynomials on both sides of the equation yields:

$$\begin{aligned} A &= a^2 \frac{da}{dt}, \\ B_1 \left(1 - \frac{6}{5} \frac{b_2}{a} \right) &= \frac{1}{2} a^3 U, \\ \frac{db_2}{dt} + 2 \frac{A}{a^3} b_2 &= 3 \frac{B_2}{a^4}, \\ \frac{db_3}{dt} + 2 \frac{A}{a^3} b_3 &= 4 \frac{B_3}{a^5} - \frac{18}{5} \frac{B_1}{a^4} b_2 \end{aligned} \quad [2.17]$$

At the surface of the bubble, the pressure must be uniform and equal to the internal gas pressure. Using Bernoulli's equation, this condition can be written as

$$gz - gR\cos\theta + \left(\frac{\partial\Phi}{\partial t}\right)_R - \frac{1}{2}(\nabla\Phi)_R^2 = \frac{k}{\rho}(\rho_g)\gamma \quad [2.18]$$

If one substitutes for Φ and R in eq. [2.18], using eqs. [2.14] - [2.15], multiplies the resulting equation in turn by each of the Legendre polynomials, and then integrates over $\cos\theta$, the orthogonality relations among the Legendre polynomials produce 4 differential equations:

$$\begin{aligned} gz + a\left(\frac{d^2a}{dt^2}\right) + \frac{3}{2}\left(\frac{da}{dt}\right)^2 - \frac{U^2}{4} + O(U^4) &= \frac{k}{\rho}(\rho_g)\gamma. \\ \frac{1}{2}\frac{d}{dt}(a^3U) - \frac{6}{5}\int_0^t \frac{B_2}{a^2} dt' + O(U^3) &= ga^3, \\ \frac{1}{a^3}\frac{dB_2}{dt} - \frac{3B_2}{a^4}\frac{da}{dt} + \frac{3}{4}U^2 - \frac{1}{a^2}\frac{d^2a}{dt^2}\int_0^t \frac{3B_2}{a^2} dt' \\ + O(U^3) &= 0, \\ \frac{1}{a^4}\frac{dB_3}{dt} - \frac{4B_3}{a^5}\frac{da}{dt} - \frac{1}{a^2}\frac{d^2a}{dt^2}\int_0^t \frac{4B_3}{a^3} dt' + O(U^3) &= 0 \end{aligned} \quad [2.19]$$

Ward (1943) has estimated the value of b_2 , which measures the departures from sphericity, in eq. [2.15], and found that it remains small until the bubble begins to contract. Near the bubble's minimum radius, b_2 becomes greater than a , making the whole calculation invalid.

III. EQUATIONS OF MOTION FOR AN ELLIPSOIDAL BUBBLE

Now, let a , b , c be the semi-axes of an ellipsoid along the x , y , z axes, respectively, for a co-ordinate system whose origin is at the centre of the ellipsoid. Let the ellipsoid be immersed in a fluid of infinite extent, and let one of the axes, say a , vary with respect to time. At any instant in time, the equation of the ellipsoid will be given by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \quad [3.1]$$

The velocity potential, Φ , for the flow about the ellipsoid is given by the solution to Laplace's equation,

$$\nabla^2\Phi = 0 \quad [3.2]$$

with appropriate boundary conditions. Since the problem obviously possesses ellipsoidal symmetry, it is most convenient to transform to ellipsoidal co-ordinates, thusly:

$$\begin{aligned}
x^2 &= \frac{(a^2+\lambda)(a^2+\mu)(a^2+\nu)}{(a^2-b^2)(a^2-c^2)}, \\
y^2 &= \frac{(b^2+\lambda)(b^2+\mu)(b^2+\nu)}{(b^2-c^2)(b^2-a^2)}, \\
z^2 &= \frac{(c^2+\lambda)(c^2+\mu)(c^2+\nu)}{(c^2-a^2)(c^2-b^2)},
\end{aligned} \tag{3.3}$$

where λ, μ, ν , are the ellipsoidal co-ordinates. The surfaces defined by $\lambda = \text{constant}$, $\mu = \text{constant}$, $\nu = \text{constant}$, are confocal quadrics. Because of the symmetry which exists in the transformation equations, eq. [3.3], one is allowed to specify which co-ordinate's constancy will yield confocal ellipsoids. Throughout this paper, then, the relation $\lambda = \text{constant}$ will be taken to describe the family of confocal ellipsoids. In ellipsoidal co-ordinates, Laplace's equation, eq. [3.2], is given by:

$$\begin{aligned}
(\mu-\nu)(k_\lambda \frac{\partial}{\partial \lambda})^2 \Phi + (\nu-\lambda)(k_\mu \frac{\partial}{\partial \mu})^2 \Phi \\
+ (\lambda-\mu)(k_\nu \frac{\partial}{\partial \nu})^2 \Phi = 0
\end{aligned} \tag{3.4}$$

where k_λ is given by

$$k_\lambda = ((a^2+\lambda)(b^2+\lambda)(c^2+\lambda))^{\frac{1}{2}} \tag{3.5}$$

and the expressions for k_μ and k_ν can be obtained from eq. [3.5] by symmetry.

Now, let α be a solution to eq. [3.4], and let another solution be given by

$$\Phi = \alpha \chi(\lambda) \tag{3.6}$$

where χ is a function of λ only. By substituting eq. [3.6] into eq. [3.4], one finds that α must have the form

$$\alpha = \alpha_\lambda f(\mu, \nu) \tag{3.7}$$

where α_λ is a function of λ , only (Milne-Thompson 1949). Using eq. [3.6] to aid in the solution of eq. [3.4], one finds that

$$\chi(\lambda) = A \int \frac{d\lambda}{\alpha_\lambda^2 k_\lambda} + B \tag{3.8}$$

where A and B are arbitrary constants. Hence, if α is a solution to eq. [3.2],

$$\Phi = \alpha \int \frac{d\lambda}{\alpha_\lambda^2 k_\lambda} \tag{3.9}$$

is also a solution. The solutions to eq. [3.4] having the form of eq. [3.9] are the ellipsoidal harmonics.

In the case of the spherical bubble, Taylor and Herring assumed that its pulsations would be described by simple radial oscillations. Since a sphere is a degenerate ellipsoid, it would be reasonable to choose as a solution to eq. [3.4] the ellipsoidal harmonic which produces analogous oscillations. Such a velocity potential is:

$$\Phi = A \int_{\lambda}^{\infty} \frac{d\lambda}{k_{\lambda}} \quad [3.10]$$

since $\alpha_{\lambda} = 1$ is a solution to eq. [3.4]. The upper limit of the integral has been chosen in order that the potential become 0 at an infinite distance from its source. The constant of integration A is determined by the boundary conditions. Now, the boundary conditions for a pulsating ellipsoid are not as obvious as those for a sphere. Nonetheless, it seems apparent that at that point on an axis which is on the surface of the ellipsoid, the normal velocity of the fluid must be equal to the rate of change with respect to time of that axis.

Hence, for an ellipsoid in which only one axis is allowed to vary, say a,

$$\left. \frac{\partial \Phi}{\partial n} \right|_{\substack{x=a \\ y=0 \\ z=0}} = -\dot{a} \quad [3.11]$$

where $\dot{a} = \frac{da}{dt}$, $\frac{\partial \Phi}{\partial n} = \nabla \Phi \cdot \hat{n}$ is the normal derivative of the potential, and \hat{n} is the unit outward normal to the ellipsoid. In ellipsoidal coordinates, eq. [3.11] can be written as:

$$\left. \frac{2}{(\mu\nu)^{\frac{1}{2}}} \frac{abc}{k_{\lambda}} \right|_{\substack{x=a \\ y=0 \\ z=0}} = \dot{a} \quad [3.12]$$

where the potential Φ is given by eq. [3.10].

Now, at $x=a$, $y=0$, $z=0$

$$\begin{aligned} \lambda &= 0 \\ \mu &= -b^2 \\ \nu &= -c^2 \end{aligned} \quad [3.13]$$

and

$$k_{\lambda} = abc \quad [3.14]$$

Hence,

$$A = \frac{1}{2} \dot{a}bc \quad [3.15]$$

and

$$\Phi = \frac{1}{2} \dot{a}bc \int_{\lambda}^{\infty} \frac{d\lambda}{k_{\lambda}} \quad [3.16]$$

The kinetic energy of the fluid around the ellipsoid is given by

$$T = -\frac{1}{2} \rho \oint \frac{\partial \Phi}{\partial n} \Phi dS \quad [3.17]$$

where the integral is carried out over a bounding ellipsoid described by eq. [3.1], and over a second outer boundary which is obtained by allowing λ in eq. [3.3] to approach infinity, essentially describing an infinitely extended ellipsoid. The contribution to eq. [3.17] from the large ellipsoid vanishes, leaving only that at the inner surface i.e. the bubble itself. The surface integral can be transformed into one over the x-y plane, thusly:

$$\oint \Phi \frac{\partial \Phi}{\partial n} dS = \iint \Phi \frac{\partial \Phi}{\partial n} \frac{1}{(\hat{n} \cdot \hat{z})} dx dy \quad [3.18]$$

where \hat{z} is the unit normal along the z axis. The unit normal, \hat{n} , to the surface of the ellipsoid is given by

$$\hat{n} = \frac{1}{\left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4}\right)^{\frac{1}{2}}} \left(\frac{x}{a^2} \hat{x} + \frac{y}{b^2} \hat{y} + \frac{z}{c^2} \hat{z}\right) \quad [3.19]$$

in the ellipsoidal co-ordinates,

$$\left(\frac{\partial \Phi}{\partial n}\right)_{\lambda=0} = \frac{1}{h_1} \left(\frac{\partial \Phi}{\partial \lambda}\right)_{\lambda=0} \quad [3.20]$$

where

$$h_1^2 = \frac{1}{4} \left(\frac{x^2}{(a^2+\lambda)^2} + \frac{y^2}{(b^2+\lambda)^2} + \frac{z^2}{(c^2+\lambda)^2} \right) \quad [3.21]$$

So, at the surface of the ellipsoid, $\lambda=0$,

$$\frac{\partial \Phi}{\partial n} = - \frac{1}{\left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4}\right)^{\frac{1}{2}}} \left(\frac{\dot{a}}{a}\right) \quad [3.22]$$

Using eqs. [3.16], [3.18], [3.19], and [3.22] and taking account of the contributions from the half surfaces above and below the x-y plane, eq. [3.17] becomes

$$T = \frac{1}{2} \frac{cb \dot{a}^2}{a} \int_0^\infty \frac{d\lambda}{k_\lambda} \int_{-a}^a \int_{b(1 - \frac{x^2}{a^2})^{\frac{1}{2}}}^{b(1 - \frac{x^2}{a^2})^{\frac{1}{2}}} \frac{c^2}{z} dx dy \quad [3.23]$$

where

$$z = c(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2})^{\frac{1}{2}} \quad [3.24]$$

Evaluating eq. [3.23] finally yields the expression for the kinetic energy, T , of the flow about an ellipsoid, described by eq. [3.1], when the semi-axis a is allowed to vary with respect to time:

$$T = \pi \rho \dot{a}^2 b^2 c^2 \int_0^\infty \frac{d\lambda}{k_\lambda} \quad [3.25]$$

Since the equations are symmetrical with respect to all three semi-axes, it is possible to obtain expressions for the kinetic energy of the flow about an ellipsoid when each of the other semi-axes are varying by means of cyclic permutations of a , b , and c , thusly:

$$T = \pi \rho a \dot{b}^2 c^2 \int_0^\infty \frac{d\lambda}{k_\lambda} \quad [3.26]$$

$$T = \pi \rho a^2 b \dot{c}^2 \int_0^\infty \frac{d\lambda}{k_\lambda} \quad [3.27]$$

where $\dot{b} = \frac{db}{dt}$ and $\dot{c} = \frac{dc}{dt}$.

Hence, by adding eqs. [3.25] - [3.27], and renormalizing, one can obtain the kinetic energy, T , for the flow around an ellipsoid when all three axes are allowed to vary in time:

$$T = \frac{\pi \rho}{3} (\dot{a}^2 b^2 c^2 + a^2 \dot{b}^2 c^2 + a^2 b^2 \dot{c}^2) \int_0^\infty \frac{d\lambda}{k_\lambda} \quad [3.28]$$

The normalisation factor 3 has been chosen so that eq. [3.28] is in agreement with the term for the kinetic energy due to radial spherical pulsations in eq. [2.1] when $a=b=c$ and $\dot{a}=\dot{b}=\dot{c}$.

As mentioned above, it is not entirely clear what sort of boundary conditions are applicable at the surface of a pulsating ellipsoid. In fact, the specification of boundary conditions when all three semi-axes are varying is equivalent to putting constraints on the interactions among \dot{a} , \dot{b} , and \dot{c} , which, in turn, is equivalent to specifying which shapes the bubble will be allowed to assume. Equation [3.28] corresponds to a velocity potential which has been so constructed that the velocity cross-terms in eq. [3.17] cancel out. Since \dot{a} , \dot{b} , and \dot{c} , are all mutually perpendicular, this seems physically reasonable. This has the effect of insisting that the movement of a point on an axis of the ellipsoid is due only to the change in length of that axis, with the changes in the lengths of the other two axes contributing nothing.

The velocity potential associated with the purely translational motion of the ellipsoid along the z axis is well known (e.g. Milne-Thompson 1949). The boundary condition is

$$-\left. \frac{\partial \Phi}{\partial n} \right)_{\lambda=0} = U \cos \theta_z \quad [3.29]$$

where, as before, $-U = \frac{dz}{dt} = \dot{z}$, θ_z is the angle between the z axis and the normal to the surface of the ellipsoid, and z is the position of the bubble centre below the pressure datum. Since

$$\cos \theta_z = \frac{1}{h_1} \left. \frac{\partial z}{\partial \lambda} \right)_{\lambda=0} \quad [3.30]$$

eq. [3.29] becomes

$$\left. \frac{\partial \Phi}{\partial \lambda} \right)_{\lambda=0} = -U \left. \frac{\partial z}{\partial \lambda} \right)_{\lambda=0} \quad [3.31]$$

The solution to eq. [3.4] which satisfies the boundary conditions in eq. [3.31] is the first order ellipsoidal harmonic given by

$$\Phi = Cz \int_{\lambda}^{\infty} \frac{d\lambda}{(c^2 + \lambda)k_{\lambda}} \quad [3.32]$$

where C is a constant of integration. Direct substitution of eq. [3.32] back into eq. [3.31] yields

$$C = -\frac{abc}{2-\alpha_0} U \quad [3.33]$$

where

$$\alpha_0 = abc \int_0^{\infty} \frac{d\lambda}{(c^2 + \lambda)k_{\lambda}} \quad [3.34]$$

The kinetic energy of the flow around a translating ellipsoid is then given by:

$$\begin{aligned} T &= -\frac{1}{2} \rho \iint \Phi \frac{\partial \Phi}{\partial n} dS \\ &= \frac{\alpha_0}{2(2-\alpha_0)} \rho U^2 \iint z \cos \theta_z dS \\ &= \frac{2\pi}{3} abc \rho \left(\frac{\alpha_0}{2-\alpha_0} \right) U^2 \end{aligned} \quad [3.35]$$

As before, the integration is carried out over a bounding ellipsoid whose semi-axes are a , b , and c , and over one whose semi-axes are allowed to extend to infinity, where the contribution from the outer ellipsoid vanishes. Hence, the kinetic energy for the flow generated by a translating ellipsoid whose semi-axes are varying in time is given by:

$$T = \frac{\pi \rho}{3} (\dot{a}^2 b^2 c^2 + a^2 \dot{b}^2 c^2 + a^2 b^2 \dot{c}^2) \int_0^\infty \frac{d\lambda}{k_\lambda} \quad [3.36]$$

$$+ \frac{2\pi}{3} abc \rho \left(\frac{\alpha_0}{2 - \alpha_0} \right) U^2$$

Equation [3.36] has no provision for the interaction of \dot{a} , \dot{b} , or \dot{c} with U . Again, this is not unexpected since eq. [3.35] should yield the kinetic energy terms in eq. [2.1], as it does in fact do, when $a=b=c$ and $\dot{a}=\dot{b}=\dot{c}$. Physically, as long as the bubble is symmetrical with respect to the x - y plane, the contributions from the top and bottom halves to a coupling of the oscillatory motion with the translational exactly cancel each other.

The energy, V_p , associated with the hydrostatic pressure around the bubble is:

$$V_p = \frac{4\pi}{3} \rho g a b c z \quad [3.37]$$

The internal energy of the bubble, $E(a,b,c)$, is given by

$$E(a,b,c) = \frac{k M \gamma a^{-(\gamma-1)} b^{-(\gamma-1)} c^{-(\gamma-1)}}{(\gamma-1) \left(\frac{4\pi}{3} \right) \gamma^{-1}} \quad [3.38]$$

Hence, the Lagrangian, L , for the flow around an ellipsoidal bubble whose semi-axes are a , b , c , at a depth z below the pressure datum, where the z axis is parallel to the ellipsoid axis c , and which is moving with a translational velocity $U = -\frac{dz}{dt}$ is given by

$$L = \frac{\pi}{3} \rho (\dot{a}^2 b^2 c^2 + a^2 \dot{b}^2 c^2 + a^2 b^2 \dot{c}^2) \int_0^\infty \frac{d\lambda}{k_\lambda} \\ + \frac{2\pi}{3} \rho a b c \left(\frac{\alpha_0}{2 - \alpha_0} \right) U^2 - \frac{4\pi}{3} \rho a b c g z \\ - \frac{k M \gamma a^{-(\gamma-1)} b^{-(\gamma-1)} c^{-(\gamma-1)}}{(\gamma-1) \left(\frac{4\pi}{3} \right) \gamma^{-1}} \quad [3.39]$$

The total energy of the bubble, $Y(t)$, at any time t is therefore given by:

$$Y(t) = \frac{\pi}{3} \rho (\dot{a}^2 b^2 c^2 + a^2 \dot{b}^2 c^2 + a^2 b^2 \dot{c}^2) \int_0^\infty \frac{d\lambda}{k_\lambda} \\ + \frac{2\pi}{3} \rho a b c \left(\frac{\alpha_0}{2 - \alpha_0} \right) U^2 + \frac{4\pi}{3} \rho a b c g z \\ + \frac{k M \gamma a^{-(\gamma-1)} b^{-(\gamma-1)} c^{-(\gamma-1)}}{(\gamma-1) \left(\frac{4\pi}{3} \right) \gamma^{-1}} \quad [3.40]$$

Hence, the equations of motion for the bubble are:

$$\begin{aligned}\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{a}} \right) - \frac{\partial L}{\partial a} &= Q_a, \\ \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{b}} \right) - \frac{\partial L}{\partial b} &= Q_b, \\ \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{c}} \right) - \frac{\partial L}{\partial c} &= Q_c, \\ \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{z}} \right) - \frac{\partial L}{\partial z} &= Q_z,\end{aligned}\tag{3.41}$$

where the Q_i are the generalised dissipative forces.

The substitution of eq. [3.39] into eq. [3.41] results in:

$$\begin{aligned}\frac{2\pi}{3} \rho b^2 c^2 I_o \ddot{a} &= - \frac{2\pi}{3} \rho \left[b^2 c^2 \frac{\partial I_o}{\partial a} \dot{a}^2 \right. \\ &+ (2c^2 b I_o + b^2 c^2 \frac{\partial I_o}{\partial b}) \dot{a} \dot{b} + (2b^2 c I_o + b^2 c^2 \frac{\partial I_o}{\partial c}) \dot{a} \dot{c} \Big] \\ &+ \frac{\pi}{3} \rho (\dot{a}^2 b^2 c^2 + a^2 \dot{b}^2 c^2 + a^2 b^2 \dot{c}^2) \frac{\partial I_o}{\partial a} \\ &+ \frac{2\pi}{3} \rho (a c^2 \dot{b}^2 + a b^2 \dot{c}^2) I_o + \frac{2\pi}{3} \rho b c f(\alpha_o) \dot{z}^2 \\ &+ \frac{2\pi}{3} \rho a b c \frac{\partial f(\alpha_o)}{\partial a} \dot{z}^2 - \frac{4\pi}{3} \rho g b c z \\ &+ (\gamma-1) K a^{-(\gamma-1)} b^{-(\gamma-1)} c^{-(\gamma-1)} + Q_a,\end{aligned}\tag{3.42}$$

$$\begin{aligned}\frac{2\pi}{3} \rho a^2 c^2 I_o \ddot{b} &= - \frac{2\pi}{3} \rho \left[a^2 c^2 \frac{\partial I_o}{\partial b} \dot{b}^2 \right. \\ &+ (2a c^2 I_o + a^2 c^2 \frac{\partial I_o}{\partial a}) \dot{a} \dot{b} + (2a^2 c I_o + a^2 c^2 \frac{\partial I_o}{\partial c}) \dot{b} \dot{c} \Big] \\ &+ \frac{\pi}{3} \rho (\dot{a}^2 b^2 c^2 + a^2 \dot{b}^2 c^2 + a^2 b^2 \dot{c}^2) \frac{\partial I_o}{\partial b} \\ &+ \frac{2\pi}{3} \rho (b c^2 \dot{a}^2 + b a^2 \dot{c}^2) I_o + \frac{2\pi}{3} \rho a c f(\alpha_o) \dot{z}^2 \\ &+ \frac{2\pi}{3} \rho a b c \frac{\partial f(\alpha_o)}{\partial b} \dot{z}^2 - \frac{4\pi}{3} \rho g a c z \\ &+ (\gamma-1) K a^{-(\gamma-1)} b^{-(\gamma-1)} c^{-(\gamma-1)} + Q_b,\end{aligned}\tag{3.43}$$

$$\frac{2\pi}{3} \rho a^2 b^2 I_o \ddot{c} = - \frac{2\pi}{3} \rho \left[a^2 b^2 \frac{\partial I_o}{\partial c} \dot{c}^2 \right.$$

$$\begin{aligned}
& + (2ab^2I_0 + a^2b^2\frac{\partial I_0}{\partial a})\dot{a}\dot{c} + (2a^2bI_0 + a^2b^2\frac{\partial I_0}{\partial b})\dot{b}\dot{c} \quad [3.44] \\
& + \frac{\pi}{3} \rho(\dot{a}^2b^2c^2 + a^2\dot{b}^2c^2 + a^2b^2\dot{c}^2) \frac{\partial I_0}{\partial c} \\
& + \frac{2\pi}{3} \rho(\dot{a}^2b^2c + a^2\dot{b}^2c)I_0 + \frac{2\pi}{3} \rho abf(\alpha_0)\dot{z}^2 \\
& + \frac{2\pi}{3} \rho abc \frac{\partial f(\alpha_0)}{\partial c} \dot{z}^2 - \frac{4\pi}{3} \rho gabz \\
& + (\gamma-1)Ka^{-(\gamma-1)}b^{-(\gamma-1)}c^{-\gamma} + Q_c,
\end{aligned}$$

$$\begin{aligned}
\ddot{z} = & - \left(\frac{\dot{a}}{a} + \frac{\dot{b}}{b} + \frac{\dot{c}}{c} \right) \dot{z} - \frac{2}{\alpha_0(2-\alpha_0)} \frac{d\alpha_0}{dt} \dot{z} \quad [3.45] \\
& - \frac{g}{f(\alpha_0)} + Q_z,
\end{aligned}$$

where

$$\begin{aligned}
\ddot{a} &= \frac{d^2a}{dt^2}, \quad \ddot{b} = \frac{d^2b}{dt^2}, \quad \ddot{c} = \frac{d^2c}{dt^2}, \\
\ddot{z} &= \frac{d^2z}{dt^2}, \quad [3.46]
\end{aligned}$$

$$I_0 = \int_0^\infty \frac{d\lambda}{k_\lambda},$$

$$f(\alpha_0) = \frac{\alpha_0}{(2-\alpha_0)},$$

$$\frac{dI_0}{dt} = \frac{\partial I_0}{\partial a} \dot{a} + \frac{\partial I_0}{\partial b} \dot{b} + \frac{\partial I_0}{\partial c} \dot{c},$$

$$\frac{df(\alpha_0)}{dt} = \frac{\partial f(\alpha_0)}{\partial a} \dot{a} + \frac{\partial f(\alpha_0)}{\partial b} \dot{b} + \frac{\partial f(\alpha_0)}{\partial c} \dot{c},$$

$$K = \frac{kM^\gamma}{(\gamma-1)\left(\frac{4\pi}{3}\right)^{\gamma-1}}$$

In order to complete the derivation of the equations of motion of the bubble, it is necessary to determine Q_a , Q_b , Q_c and Q_z , the generalised dissipative forces associated with the radiation of sound by the bubble. Now, at distances large compared with the scale of the bubble, the form of the velocity potential, Φ , in the fluid will be identical with that of a spherical bubble. Hence,

$$\Phi = -\frac{C}{r} + \vec{A} \cdot \vec{\nabla}(1/r) \quad [3.47]$$

where C and $\hat{\lambda}$ are constants which depend only on the time t . r is the distance to the field point from an origin located somewhere within the bubble.

Following the development given by Landau and Lifshitz (1966), in the wave zone,

$$\Phi = \frac{C(t')}{r} + \vec{V} \cdot (A(t') \frac{\cos \theta}{r} \hat{r}) \quad [3.48]$$

where θ is the angle between the direction of the translational motion and r , and the retarded time t' is given by

$$t' = t - \frac{r}{C_S} \quad [3.49]$$

where C_S is, as before, the velocity of sound in water. The velocity, \vec{V} , of the water in the wave zone must therefore be given by:

$$\begin{aligned} \vec{V} &= -\vec{\nabla} \Phi \\ &\approx \left(\frac{1}{C_S r} \frac{\partial C(t)}{\partial t} - \frac{\cos \theta}{C_S^2 r} \frac{\partial^2 A(t)}{\partial t^2} \right) \hat{r} \end{aligned} \quad [3.50]$$

+...

where terms of higher negative order in r have been neglected. The total energy emitted as sonic radiation per unit time, $\frac{dE}{dt}$ is then:

$$\begin{aligned} \frac{dE}{dt} &= -\rho C_S \iint (\vec{V} \cdot \vec{V}) dS \\ &= -\frac{4\pi\rho}{C_S} \left(\frac{\partial C(t)}{\partial t} \right)^2 - \frac{4\pi\rho}{3C_S^3} \left(\frac{\partial^2 A(t)}{\partial t^2} \right)^2 \end{aligned} \quad [3.51]$$

where the integral has been taken over a sphere of radius r (Landau and Lifshitz 1966).

To a good approximation, the term in eq. [3.51] proportional to C_S^{-3} can be neglected for low translational velocities, since it will be 2 orders of magnitude smaller than that proportional to C_S^{-1} .

Now, the volume $4\pi C$ of fluid which flows through the surface over which the integral in eq. [3.51] is taken must be equal to the rate of change with respect to time of the volume, \hat{V} , of the bubble. Thus,

$$\begin{aligned} C &= \frac{1}{4\pi} \hat{V} \\ &= \frac{1}{3} (\dot{a}bc + a\dot{b}c + ab\dot{c}) \end{aligned} \quad [3.52]$$

and so

$$\begin{aligned} \frac{dE}{dt} = & -\frac{4\pi\rho}{9C_S} [\ddot{a}bc + \ddot{a}\dot{b}\dot{c} + a\ddot{b}\dot{c} \\ & + \dot{a}(\dot{b}\dot{c} + \dot{b}\ddot{c}) \\ & + \dot{b}(\dot{a}\dot{c} + \dot{a}\ddot{c}) \\ & + \dot{c}(\dot{a}\dot{b} + \dot{a}\ddot{b})]^2 \end{aligned} \quad [3.53]$$

Now, evidently,

$$\frac{dE}{dt} = \dot{a}Q_a + \dot{b}Q_b + \dot{c}Q_c + \dot{z}Q_z \quad [3.54]$$

Since all of the dependence on the translational velocity in eq. [3.51] was contained in the term proportional to C_S^{-3} , it follows that, to the same approximation, $Q_z = 0$ in eq. [3.54]. Since, in the absence of any translational motion, there exists nothing to distinguish one axis of the ellipsoid from another, it follows that one should be able to obtain the other two Q_i from one by cyclic permutation of the axes a , b , and c . The only grouping of the terms in eq. [3.53] which is invariant under cyclic permutation of the axes is given by

$$\begin{aligned} \frac{dE}{dt} = & -\frac{4\pi}{9C_S} \rho \left[F(\ddot{a}) \sum_{i=1}^3 F(\ddot{a}_i) \right. \\ & \left. + F(\ddot{b}) \sum_{i=1}^3 F(\ddot{a}_i) + F(\ddot{c}) \sum_{i=1}^3 F(\ddot{a}_i) \right] \end{aligned} \quad [3.55]$$

where

$$\begin{aligned} F(\ddot{a}_i) = & \ddot{a}_i a_j a_k + \dot{a}_i \dot{a}_j a_k \\ & + \dot{a}_i \dot{a}_k a_j, \quad i \neq j \neq k \end{aligned} \quad [3.56]$$

and $a_1 = a$, $a_2 = b$, $a_3 = c$.

Hence, by comparison of eq. [3.55] with eq. [3.54],

$$Q_{a_i} = -\frac{4\pi}{9C_S} \rho \frac{F(\ddot{a}_i)}{\dot{a}_i} \sum_{j=1}^3 F(\ddot{a}_j) \quad [3.57]$$

where, as before, i, j, k , are successively equal to 1, 2, 3. Now, as was the case for the spherical bubble, the terms in eq. [3.57] which depend upon the products of the pulsational accelerations with themselves or with the pulsational velocities are analogous to the radiation reaction terms in electromagnetic theory. This suggests that such terms

may be ignorable, at least in a first approximation. At this stage, in the absence of any perturbing force in the x-y plane, it is possible to allow $\dot{b} = \dot{a}$. This allows one to drop eq. [3.42] as an equation of motion, and replace \dot{b} and b by \dot{a} and a in eqs. [3.43] - [3.45] and eq. [3.57]. This, of course, specialises the equations of motion to those of a spheroidal bubble. Henceforth, throughout this paper, the case of the spheroidal bubble will be treated exclusively.

To sum up: eqs. [3.42] - [3.45] are equations of motion describing a pulsating ellipsoidal bubble undergoing translational motion. When $Q_a = Q_b = Q_z = 0$, the equations neglect any sort of energy loss. It has been shown that the energy loss from the translational motion of the bubble can be expected to be negligible with respect to that from the pulsational motion and so Q_z was set to 0 in eq. [3.45]. It was further shown that

$$Q_{a_i} = - \frac{4\pi}{9C_S} \rho \frac{F(\ddot{a}_i)}{\dot{a}_i} \sum_{j=1}^3 F(\ddot{a}_j) \quad [3.58]$$

where

$$F(\ddot{a}_i) = \ddot{a}_i \ddot{a}_j a_k + \ddot{a}_i \ddot{a}_k a_j + \Delta F(\ddot{a}_i), \quad i \neq j \neq k \quad [3.59]$$

When one wishes to ignore the effects of radiation reaction,

$$\Delta F(\ddot{a}_i) = 0 \quad [3.60]$$

and

$$\Delta F(\ddot{a}_i) = \ddot{a}_i \ddot{a}_j a_k, \quad i \neq j \neq k \quad [3.61]$$

when one wishes to include them.

IV. NUMERICAL METHODS OF SOLUTION

Before one attempts numerical solutions of the equations of motion, eqs. [3.43] - [3.45], it is useful to make them non-dimensional. Thus, the substitution of

$$\begin{aligned} a &= a^* L, \\ c &= c^* L, \\ z &= z^* L, \\ t &= t^* T \end{aligned} \quad [4.1]$$

into the equations of motion used, where

$$L = \left(\frac{Y_0}{g\rho} \right)^{\frac{1}{2}} \quad [4.2]$$

$$T = \sqrt{\frac{L}{g}}$$

yields a dimensionless form of the equations of motion. As before, Y_0 is given by eq. [2.3] and g and ρ are, respectively, the gravitational acceleration and the density of water. These particular scaling factors in eq. [4.2] were originally used by Taylor (1942).

Since all of the equations of motion have the unfortunate property of singularity at the origin, it is necessary to begin the integration with a series solution. Taylor (1942) suggested that the initial solutions to the dimensionless forms of eqs. [2.1] - [2.2] be

$$a^* = \left(\frac{t^*}{1.0025} \right)^{2/5},$$

$$\dot{z}^* = -\left(\frac{10}{11} \right) t^*, \quad [4.3]$$

$$z^* = z_0 - \left(\frac{5}{11} \right) t^{*2},$$

where z_0^* is the initial dimensionless depth below the pressure datum, for values of t^* near zero. Since the bubble can be expected to be spherical initially, the values for a^* , \dot{z}^* , and z^* from eq. [4.3] were used to begin the integration at time t_0^* , with the additional requirement that $a^* = c^*$. One also needs initial values for the rates of change, \dot{a}^* and \dot{c}^* , of the semi-axes. These were estimated by assuming that the bubble would be initially spherical and substituting into the dimensionless form of eq. [2.1] to find \dot{a}^* . Hence, the initial values for \dot{a}^* and \dot{c}^* are given by

$$\dot{a}^{*2} = \left(1 - \frac{E^*(a^*)}{Y_0} \right) \frac{1}{2\pi a^*} - \left(\frac{\dot{z}^*}{6} \right)^2 - \frac{2}{3} z^*, \quad [4.4]$$

$$\dot{a}^* = \dot{c}^*,$$

where

$$E^*(a^*) = \frac{kM\gamma a^{*-3(\gamma-1)}}{(\gamma-1) \left(\frac{4\pi}{3} \right)^{\gamma-1} L^{3(\gamma-1)}}, \quad [4.5]$$

Y_0 is the initial total energy, as given by eq. [2.3], and all other variables are as previously defined.

Another numerical difficulty concerns the evaluation of the terms $\frac{\partial I_0}{\partial a}$, $\frac{\partial I_0}{\partial c}$, $\frac{\partial \alpha_0}{\partial a}$, $\frac{\partial \alpha_0}{\partial c}$. Now, evidently,

$$\frac{\partial I_0}{\partial a} = -a \int_0^{\infty} \frac{d\lambda}{(a^2+\lambda)^{3/2} (b^2+\lambda)^{1/2} (c^2+\lambda)^{1/2}},$$

$$\frac{\partial I_0}{\partial c} = -c \int_0^{\infty} \frac{d\lambda}{(a^2+\lambda)^{1/2} (b^2+\lambda)^{1/2} (c^2+\lambda)^{3/2}}, \quad [4.6]$$

$$\frac{\partial \alpha_0}{\partial c} = bc I_1 + abc \frac{\partial I_1}{\partial a}, \quad [4.7]$$

$$\frac{\partial \alpha_0}{\partial c} = ab I_1 + abc \frac{\partial I_1}{\partial c}$$

where

$$I_1 = \int_0^{\infty} \frac{d\lambda}{(a^2+\lambda)^{1/2} (b^2+\lambda)^{1/2} (c^2+\lambda)^{3/2}} \quad [4.8]$$

I_0 and I_1 were evaluated with the IMSL double precision subroutines MMLINF and MMLIND, which compute incomplete elliptic integrals of the first and second kind, respectively. The partial derivatives of I_0 , eq. [4.6], are, in fact, incomplete elliptic integrals of the second kind, and can be evaluated with MMLIND. The evaluation of the second terms in eq. [4.7] presented considerable difficulty. In point of fact, no commercial routine capable of evaluating eqs. [4.7] seems to exist, and the difficulties involved in the composition of one ab nihilo are formidable. Accordingly, as a stopgap, the terms in eqs. [4.7] which involve the derivative of an incomplete elliptic integral of the second kind were evaluated by holding one of a or c constant, and varying the other at each step in the integration, thusly:

$$\frac{\partial I_1}{\partial a} = \frac{1}{2\Delta a} (I_1(a+\Delta a, c) - (I_1(a-\Delta a, c))), \quad [4.9]$$

$$\frac{\partial I_1}{\partial c} = \frac{1}{2\Delta c} (I_1(a, c+\Delta c) - I_1(a, c-\Delta c))$$

where MMLIND was used to evaluate I_1 . Since Δa and Δc can be made as small as desired, theoretically eq. [4.9] can be made to approximate the true value of the derivative as closely as desired; however, the practical constraints of computational time, machine accuracy, and the accuracy of the IMSL subroutines do place limits on the size of Δa and Δc .

The actual integrations were carried out using a 4 point Runge-Kutta algorithm incorporating automatic error controls. Some numerical difficulties with this method were encountered when the radiation reaction was incorporated using eq. [3.61]. Near the minima of a and c , \dot{a} and \dot{c} become small, and change sign as well. Because of their dependence upon \ddot{a}^{-1} and \ddot{c}^{-1} , the values of Q_a and Q_c can oscillate rapidly, adversely affecting the convergence of the integration. This difficulty was circumvented by the use of a series approximation in which \ddot{a}^{-1} and \ddot{c}^{-1} were replaced by averaged pulsational velocities. By using eq. [3.40], it is possible to write

$$\overline{a^{*4}} (\overline{\dot{a}^{*2}}) I_o^* = \left(\frac{Y(t) - E^*(a^*)}{Y_o} \right) \frac{3}{\pi} \quad [4.10]$$

$$-2a^* b^* c^* \left(\frac{\alpha_o}{2 - \alpha_o} \right) \dot{z}^{*2} - 4a^* b^* c^* z^*$$

where

$$\begin{aligned} \overline{a^{*4}} &= \frac{1}{8\pi} (a+b+c)^4 \\ \overline{\dot{a}^{*2}} &= \frac{1}{\overline{a^{*4}}} (\dot{a}^{*2} b^{*2} c^{*2} + a^{*2} \dot{b}^{*2} c^{*2} \\ &\quad + a^{*2} b^{*2} \dot{c}^{*2}) \end{aligned} \quad [4.11]$$

and $Y(t)$ is the energy of the bubble at any time t , given by eq. [3.40].

If one defines:

$$\begin{aligned} \alpha &= \left(\frac{Y(t) - E^*(a^*)}{Y_o} \right) \frac{3}{\pi} \overline{a^{*4}} I_o^* \\ \beta &= - \frac{2a^* b^* c^*}{(\overline{a^{*4}}) I_o^*} \left(\frac{\alpha_o}{2 - \alpha_o} \right) \dot{z}^{*2} - \frac{4a^* b^* c^* z^*}{(\overline{a^{*4}}) I_o^*} \end{aligned} \quad [4.12]$$

then,

$$(\overline{\dot{a}^{*2}})^{-\frac{1}{2}} = \frac{\pm 1}{\alpha^{\frac{1}{2}}} \left(1 + \frac{1}{2} \frac{\beta}{\alpha} + \frac{3}{8} \left(\frac{\beta}{\alpha} \right)^2 t \dots \right) \quad [4.13]$$

where the positive value is taken while a particular axis is expanding and the negative while it is contracting. By expanding the dimensionless form of eq. [3.58] when $\Delta F(a_i)$ is given by eq. [3.61], and substituting $(\overline{\dot{a}^{*2}})^{-\frac{1}{2}}$ for $(\dot{a}^*)^{-1}$ and $(\dot{c}^*)^{-1}$, values for the dissipative function incorporating averaged radiation reaction terms were obtained.

Estimates for a and c were arrived at by the substitution of the current values of a , c , \dot{a} , \dot{c} , z , and \dot{z} into eqs. [3.43] and [3.44] with Q_a and Q_c set to zero. Those values for a and c were substituted back into eq. [3.58], to obtain new values for Q_a and Q_c which were in turn used in eqs. [3.43] and [3.44] to obtain new estimates for a and c .

V. NUMERICAL RESULTS AND ANALYSIS

Figures 2-11 show the results of computations for a bubble produced by the detonation of 2.1136 kg. of TNT 6.1 metres below the surface, using eqs. [3.43] - [3.46] for a spheroidal bubble. Those curves associated with a spheroidal bubble and labelled ' $\Delta Y = 0$ ' were calculated under the assumption of no energy loss; that is, $Q_a = Q_c = Q_z = 0$ in eqs. [3.42] - [3.46]. The curves labelled ' $\Delta Y \neq 0$, $\Delta F = 0$ ', were

calculated incorporating radiative energy loss, but not radiation reaction terms; that is, eqs. [3.58] - [3.60] were used to define Q_a and Q_c . In Figs. 2-5, a and c are the semi-axes of the spheroidal bubble, where a is the semi-axis in the plane normal to the bubble's upward motion, and c the semi-axis in the plane parallel to the bubble's upward motion.

Taylor (1942) considered the same case, using eqs. [2.1] - [2.2] for a spherical bubble in the absence of any energy loss from any source. In Figs. 2-11, the curves labelled with ' $\Delta Y = 0$ ' were obtained by solving eqs. [2.10] - [2.11], with $Q_a = Q_z = 0$, which are equivalent to Taylor's equations for a spherical bubble. The curves labelled ' $\Delta Y \neq 0$, $\Delta Q_a = 0$ ' result from the incorporation of radiative energy loss, neglecting radiation reaction, into eqs. [2.10] - [2.11]; that is; Q_a was given by eq. [2.12], with $\Delta Q_a = 0$. The curves labelled ' $\Delta Y \neq 0$, $\Delta Q_a \neq 0$ ' incorporate energy loss including radiation reaction; that is; Q_a and ΔQ_a were given by eqs. [2.12] - [2.13]. In Figs. 2-5, 'radius' refers to the spherical bubble radius as calculated in eqs. [2.10] - [2.11].

The curves labelled 'spherical' in Figs. 6-11 are the upward velocities and heights above the original detonation point, obtained by solving eqs. [2.10] - [2.11] for a spherical bubble, under different assumptions about the nature of the energy loss. The curves labelled 'spheroidal' are the same quantities obtained from the solution of eqs. [3.43] - [3.46] for a spheroidal bubble.

Taylor (1943) presented photographs showing the behaviour of bubbles generated by electrical discharges in oil, which are here reproduced in Figs. 12-13. These show that a bubble in the early stages of its motion is very nearly spherical, but that near its minimum volume, it becomes approximately disc shaped, with its longest dimension lying in the plane normal to the direction of its upward motion. Near the second maximum, the bubble is highly non-spherical, and, in fact, seems to be attempting to fission, exhibiting an extremely large bulge on its upper surface and a flat lower surface. After the bubble has passed through its second maximum, it becomes mushroom-shaped and actually does bifurcate at its second minimum. The two halves rejoin later to form a distorted disc.

As one can see from Fig. 2, a spheroidal bubble reproduces the salient features of the observed behaviour, at least qualitatively. a and c were very nearly equal to each other, as well as to the spherical radius, at the first maximum. Near the first minimum, the bubble became more obviously spheroidal, with the ratio a/c assuming a value of 2.34. After the first minimum, the qualitative agreement between the spheroidal bubble model and Taylor's photographs was less pronounced. This was hardly surprising, given that the equations of motion constrain the

possible shapes of the bubble to spheroids. However, the model does at least predict that the bubble would not return to a spherical shape. In addition, the spheroidal model also retains one of the successful features of the spherical model, that of the prediction of the period of the bubble's oscillation. The period of the first oscillation in Fig. 2 (which was defined for both the spherical and spheroidal bubble to be the time after a maximum at which the first derivative of the volume changed sign) was .26 seconds for both models. It should also be noted that the time at which the volume was a minimum did not coincide with the minima of either a or c. In Fig. 2, a reached its first minimum a full .01 seconds before c did.

Figures 3-4 show the effects of incorporating radiative energy loss into the spheroidal bubble model, and compare the results with the analogous case for a spherical bubble. When radiative energy loss was included in the calculations, the second minima of a and c occurred slightly earlier, and was more nearly coincident. Figure 5 compares the semi-axes for a spheroidal bubble under various assumptions about the form of the energy loss. It is interesting that the most noticeable difference among the calculations was in the value of c.

The outstanding failure of the spherical model for the bubble is its prediction of a too rapid rate of rise when the bubble's volume is a minimum. As Figs. 6-10 show, the maximum upward velocity predicted by the spheroidal model was less than that predicted by the spherical model by a factor of 2, and consequently the distance travelled from the site of the explosion was decreased by about the same amount. This diminuation of the upward translational velocity of the bubble near its minimum volume accounted for the differences between the periods of the spherical and spheroidal models after the first minimum in Figs. 2-4. Because the spheroidal bubble was deeper than the spherical one, the hydrostatic pressure was greater, making the period shorter. This is also the reason that the curves in Figs. 2-11 associated with the spherical bubble model terminated .1 seconds before those of the spheroidal model. Because of the spherical bubble's greater upward velocities, it reached the surface before the spheroidal one.

Figure 11 shows the energy possessed by the bubble as a function of time for both the spherical and spheroidal models. Since the initial solutions to the equations of motion for the spheroidal bubble were calculated by assuming it to have been initially spherical, it is not surprising that, in the absence of dissipation, the energies predicted by the two models were found to be constant and equal. The energy losses predicted by the spheroidal model without radiation reaction were found to be in close agreement with those of the spherical model without radiation reaction. After the first minimum volume was passed, the spheroidal model predicted a slightly greater energy loss than the spherical, until the spheroidal bubble passed through a second minimum volume.

The most dramatic difference between the spheroidal and spherical models is in the role of radiation reaction. As can be seen in Fig. 11, the inclusion of radiation reaction in the spherical bubble model decreased the amount of energy radiated. For the spheroidal bubble, this was the case only until the first minimum volume was passed. After that time, the effect of including radiation reaction in the calculation was to increase the radiative energy loss, compared both to that of the spheroidal model without radiation reaction and to that of either spherical model. Moreover, since even without the inclusion of radiation reaction, the spheroidal model yielded a greater energy loss than the analogous spherical case, it seems that this was not an artefact of the approximations used in the computation of the radiation reaction terms. Whether the magnitude of the increase in the energy loss which occurred when radiation reaction terms were added to the equations would be as large as that indicated by Fig. 11 is rather more uncertain. The close correspondence between the predicted energy losses for the spherical and spheroidal models prior to the first minimum suggests that the calculation was valid, at least in the regime in which the bubble was nearly spherical. However, as mentioned above, after the first minimum, the actual shape of the bubble is not really spheroidal. Consequently, the actual radiation loss may be quite different to that calculated for a spheroidal bubble. However, since the bubble is even less spherical than it is spheroidal, on balance, it seems probable that the predictions of the spheroidal model were more accurate than those of the spherical.

Hicks (1972) solved the equations of motion for a spherical bubble, with the addition of a hydrodynamic drag term, for 227.27 kg. of TNT at a depth of 45.73 metres below the surface. (That is, Hicks took eqs. [2.11] as his equations of motion, with Q_z as defined by eq. [2.8] and $Q_a = 0$). It has been observed that a bubble from an explosion with these characteristics rises approximately 3.35 metres from the location of the explosion in the time taken to reach its first minimum. Hicks found that where drag is the only source of dissipation, a drag coefficient of $C_D = 2.25$ had to be introduced into the equations of motion for a spherical bubble in order to reproduce this behaviour. In this work, for the initial solutions chosen, it was found that a drag coefficient of $C_D = 1.85$ brought the predicted rise of a spherical bubble into better agreement with observation. When radiative dissipation was also included, a drag coefficient of $C_D = 1.6$ yielded better agreement with observation.

In Figs. 14-23, the curves labelled 'spherical' were obtained by taking eqs. [2.10] - [2.11] as the equations of motion for a spherical bubble produced by 227.27 kg. of TNT detonated 45.73 metres below the surface. All of the symbols in the legend for these figures have the same meaning as in Figs. 2-11. It should also be understood that, for the spherical bubble, the effects of drag were ignored unless a value of the drag coefficient C_D is given in the legend. When drag was considered, F_D was given by eq. [2.8]. It should be emphasized that a drag term was incorporated only into the equations of motion for a spherical bubble and never into eqs. [3.43] - [3.46], the equations of motion for a spheroidal bubble.

The behaviour of the spheroidal bubble in Figs. 14-17 was very similar to that in Figs. 2-5. One difference was the close agreement among all four curves for the period of the first bubble oscillation. As well, for the second oscillation, the period predicted by the spheroidal model was closer to that of the spherical model with drag than that which was predicted by the spherical model without drag. This applied in the cases for which radiative energy loss was considered as well as those cases for which it was not. In Figs. 18-21, it can be seen that the peak velocities predicted by the spheroidal models, both with and without radiative energy loss, were in agreement with those predicted by the spherical model with drag, most notably at the first minimum.

It is Fig. 22, however, which demonstrates the accuracy of the spheroidal model. As can be seen, the height above the site of the original explosion predicted by the spheroidal model was in good agreement with that predicted by the spherical model with drag, while that predicted by the spherical model in the absence of drag disagreed with that predicted by the spherical model with drag. Because the drag coefficient used with the spherical model was chosen specifically to force the calculated height to agree with observation, the agreement of the spheroidal model with the spherical model in this case constitutes a verification of the spheroidal model. The agreement was not as good at the second minimum, especially for the spheroidal bubble without energy loss. However, once radiative energy loss was added to the spheroidal model, the curves exhibited close agreement with those for the spherical model with drag.

Figure 23 compares the predicted energy losses of the spherical and spheroidal models. Once again, it can be seen that the inclusion of radiation reaction terms increased the predicted energy loss for a spheroidal bubble, in contrast to the spherical model, for which the addition of reaction terms decreased the energy loss. It is interesting, however, the energy remaining to the spheroidal bubble was greater than that remaining to the spherical bubble when losses from both drag and radiation were included.

VI. CONCLUDING REMARKS

In this work, a Lagrangian for an oscillating ellipsoidal bubble which is also undergoing translational motion has been derived, and equations of motion obtained from it. An expression for the generalised dissipative forces caused by the radiation of sound by the bubble was also found, and incorporated into the equations of motion. This equation was specialised to the case of a spheroidal bubble, and the equations of motion solved for some different charge masses and depths, both with and without the effects of radiation of sound having been included.

By comparison with the results obtained from Taylor's spherical bubble model, it was shown that the spheroidal model retained the successful features of the spherical model, notably the prediction of the bubble's oscillatory period, and reproduced, at least qualitatively, notable features of a real bubble's behaviour, such its flattening near its first minimum, and a slower rise time than that predicted by the spherical model. After the first minimum, the bubble's shape was not as well modelled by a spheroid; however, the results obtained at those times were still superior to those from the spherical model, particularly with respect to the rise time. For the case cited by Hicks (1972), the spheroidal bubble model produced results which were in very close agreement with experimental data. The spherical model was capable of similar agreement only for a limited time and only with the addition of a drag term. The spheroidal model's advantage arises from its reproduction of the height above the explosion site naturally, without the addition of a drag term which must be determined from experimental data for each case. Incidentally, because the spheroidal model ignored hydrodynamic drag completely, and still reproduced the observed behaviour of the bubble, it seems likely that drag is relatively unimportant in determining the bubble's motion; it appears, rather, that the shape of the bubble is the single most important factor.

It appears that the spheroidal model predicted a greater loss of energy in the bubble through the radiation of sound than the spherical model. This was found to be most significant when radiation reaction terms were included in the radiative dissipation function. In contrast to the spherical model, the effect of including radiation reaction terms was to increase the energy loss.

VII. BIBLIOGRAPHY

Cole, R.H. 1948, Underwater Explosions, Princeton University Press, Princeton

Heaton, K.C. 1984, in Transactions of the 2nd Army Conference on Applied Mathematics and Computing, 535, ARO Report 85-1, U.S. Army Research Office, UNCLASSIFIED

Herring, C. 1942, in Underwater Explosion Research, Vol. II, 35 Office of Naval Research, Dept. of the Navy, Washington, D.C., 1950, UNCLASSIFIED

Hicks, A.N. 1972, The Theory of Explosion Induced Whipping Ship Motions, Report no. NCRE/R579, Naval Construction Research Establishment, St. Leonard's Hill, Dunfermline, Fife UNCLASSIFIED

Holt, R.A. 1977, Annual Review of Fluid Mechanics, 9, 187 Pao Alto, California

Landau, L.D. and Lifshitz, E.M. 1966, Fluid Mechanics, Addison-Wesley Inc., Don Mills, Ontario

Milne-Thomson, L.M. 1949, Theoretical Hydrodynamics, Macmillan and Co. Ltd., St. Martin's St., London

Penney, W.G., and Price, A.T. 1942, in Underwater Explosion Research, Vol. II, 145, Office of Naval Research, Dept. of the Navy, Washington, D.C., 1950, UNCLASSIFIED

Shiffman, M. and Friedman, B. 1944, in Underwater Explosion Research, Vol. II, 245, Office of Naval Research, Dept. of the Navy, Washington, D.C., 1950, UNCLASSIFIED

Taylor, Sir G.I. 1942, in The Scientific Paperes of Sir Geoffrey Ingram Taylor, Vol. III, 320, Cambridge University Press, Cambridge 1963

_____, 1943, in The Scientific Papers of Sir Geoffrey Ingram Taylor, Vol. III, 337, Cambridge University Press, Cambridge 1963

Ward, A.B. 1943, Undex 20, cited in Cole (1948)

Figure 1

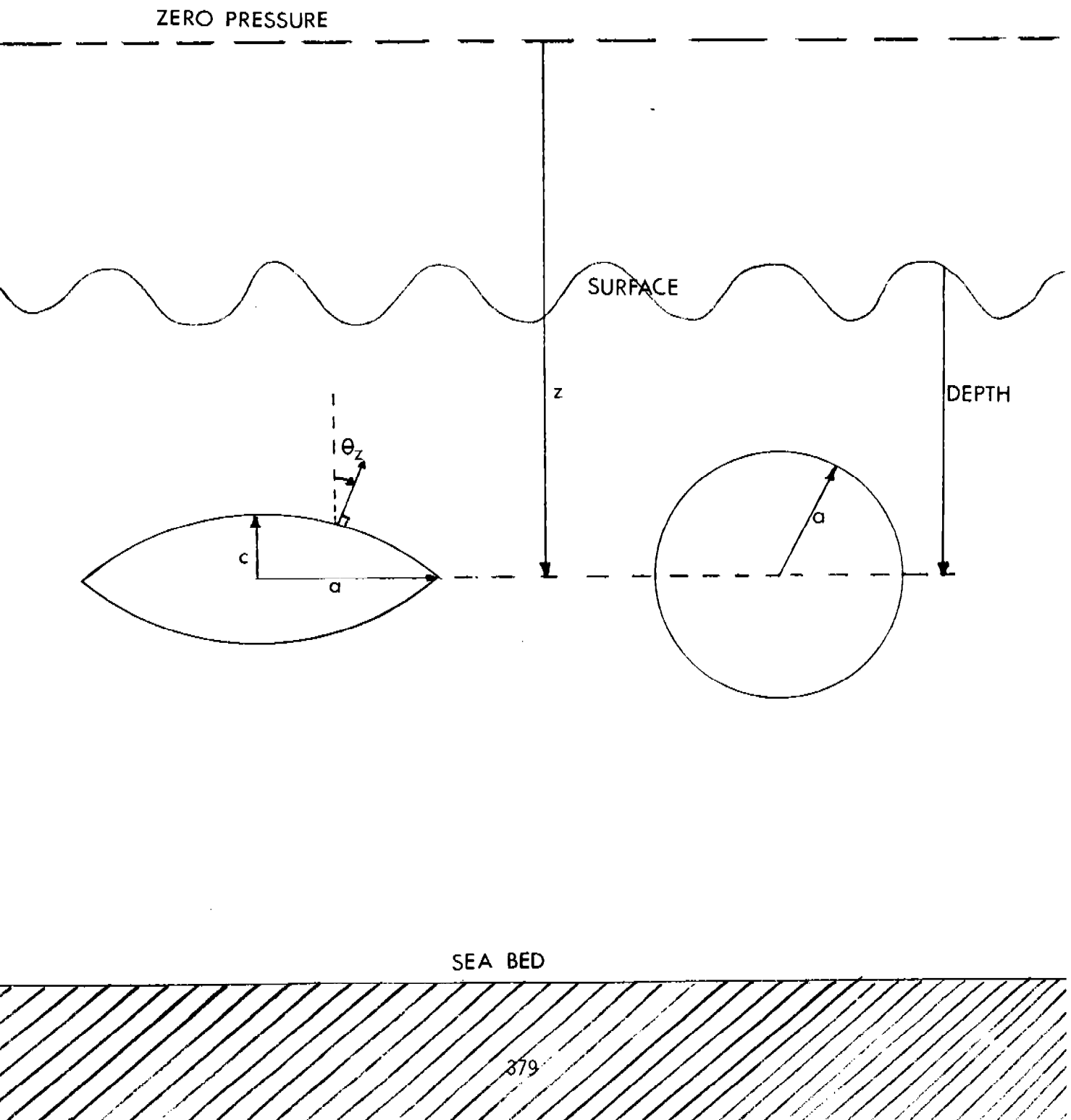


FIGURE 2

SEMI-AXES OF BUBBLE

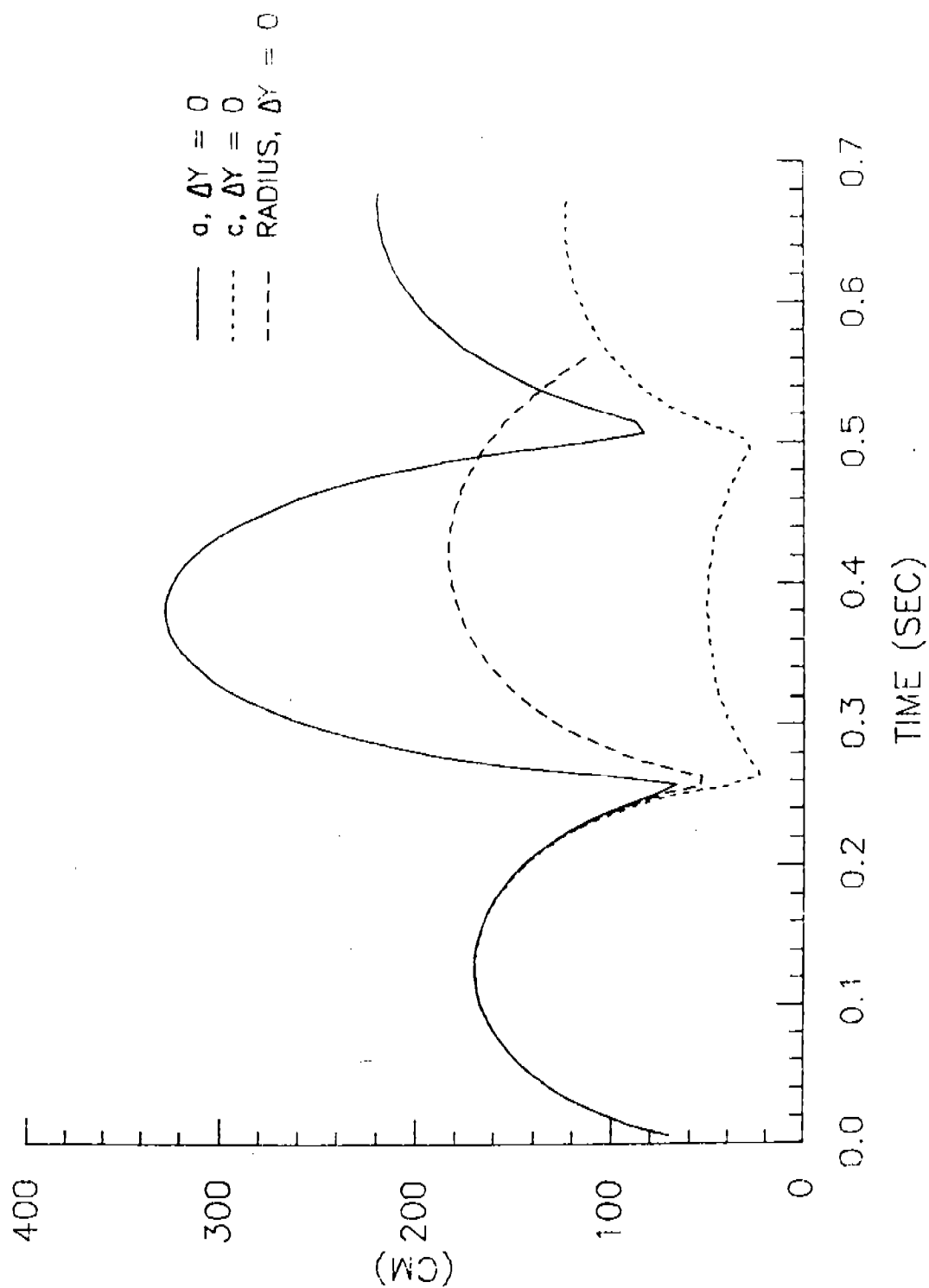


FIGURE 3

SEMI-AXES OF BUBBLE

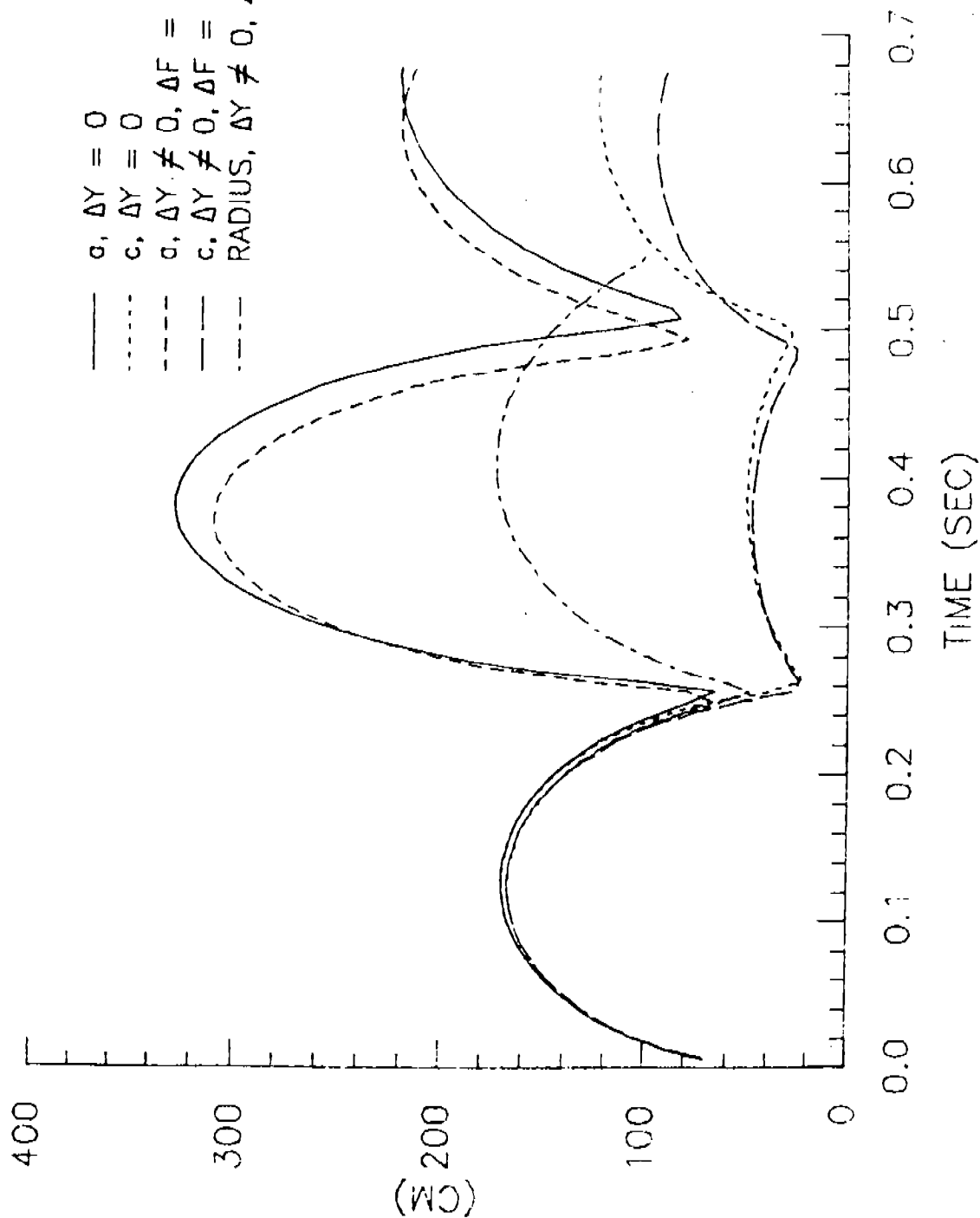


FIGURE 4

SEMI-AXES OF BUBBLE

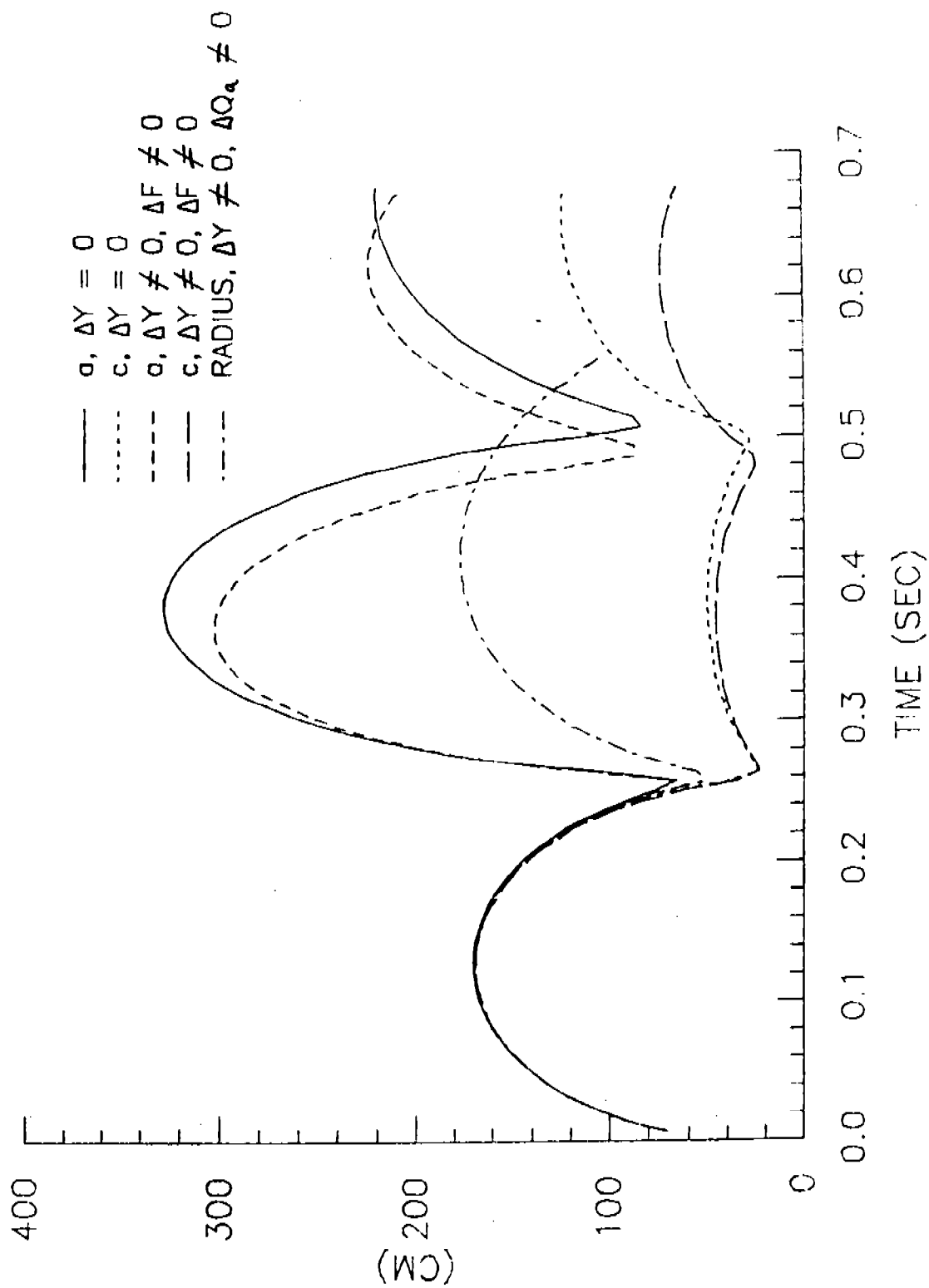


FIGURE 5

SEMI-AXES OF BUBBLE

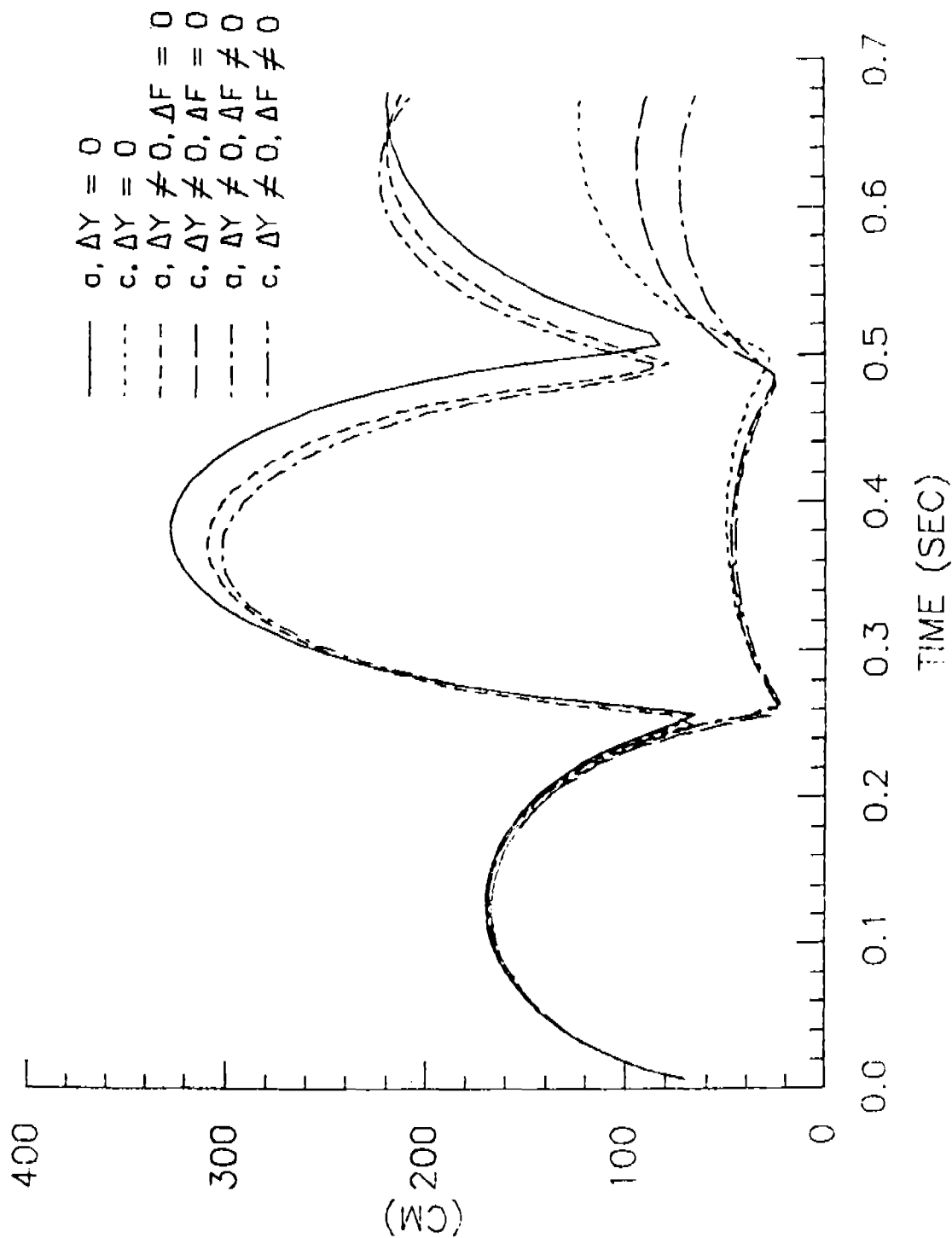


FIGURE 6

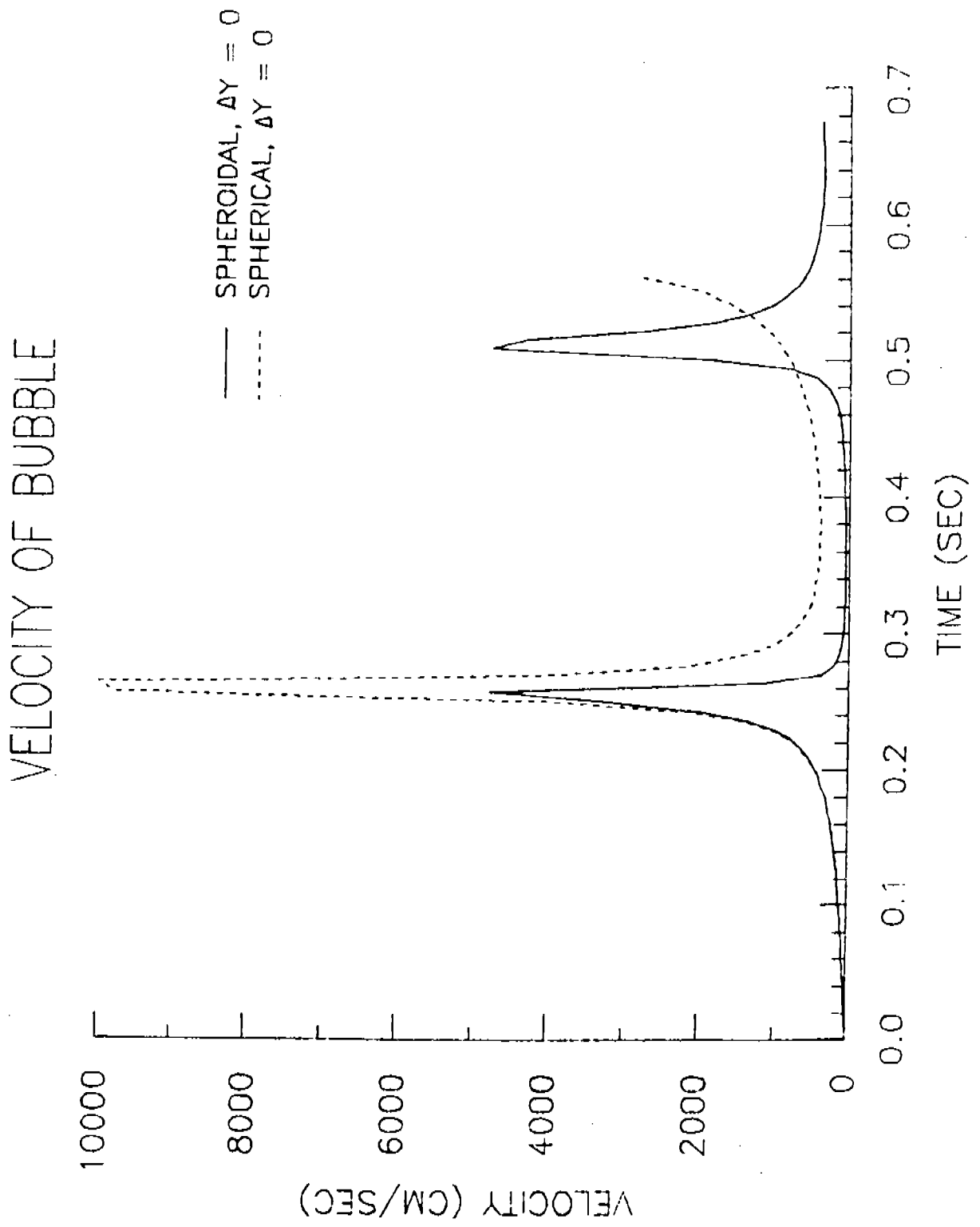


FIGURE 7

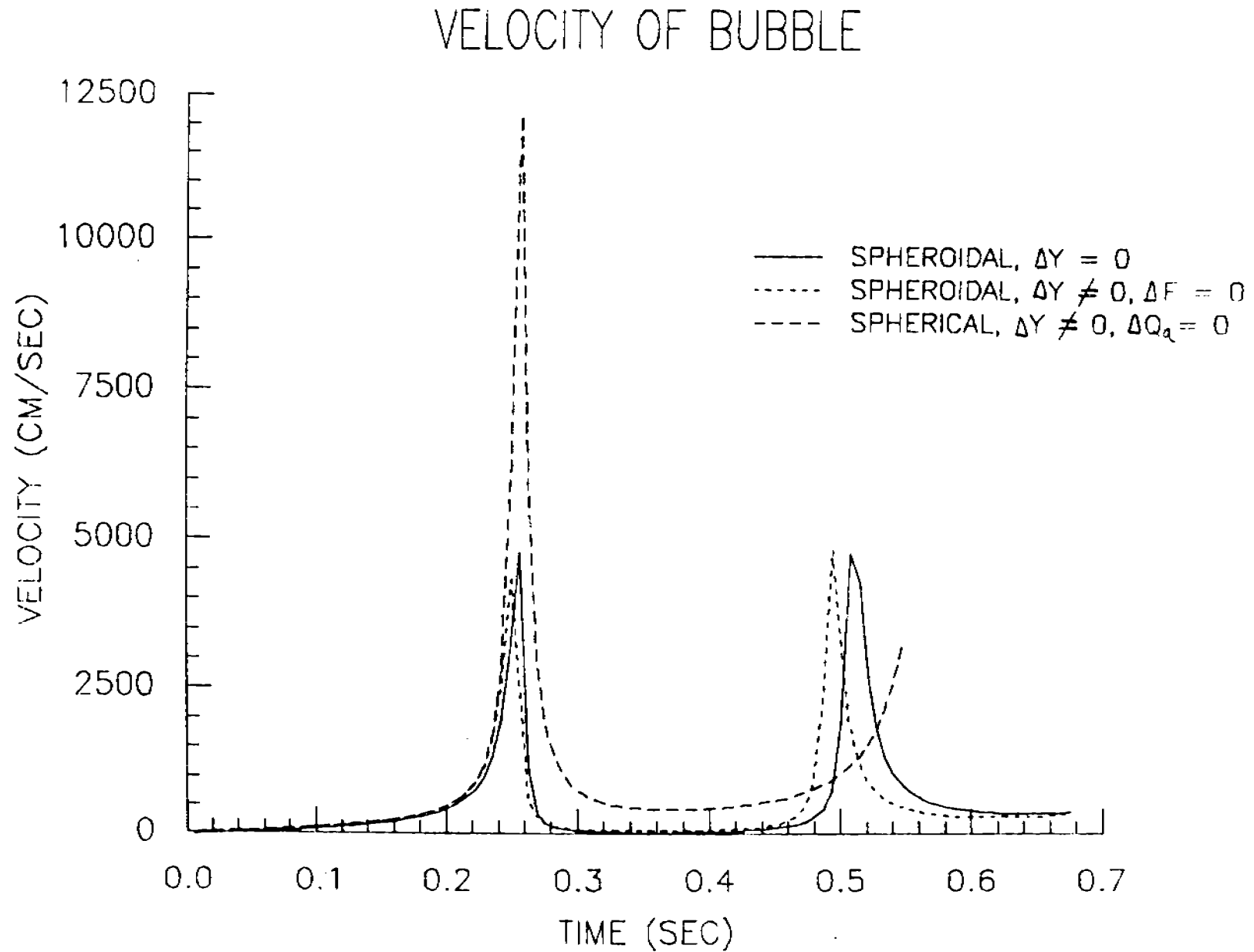


FIGURE 8

VELOCITY OF BUBBLE

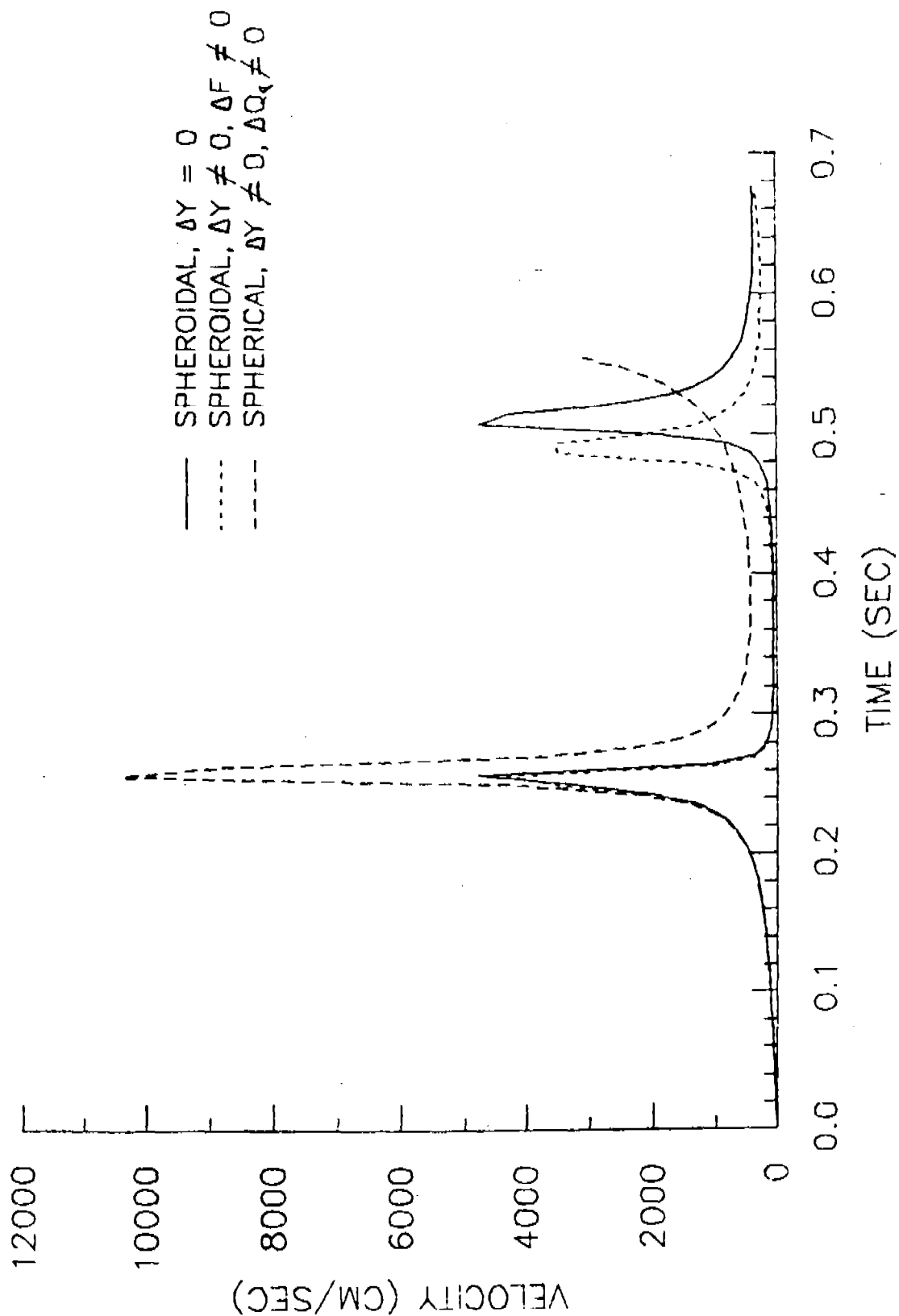


FIGURE 9

VELOCITY OF SPHEROIDAL BUBBLE

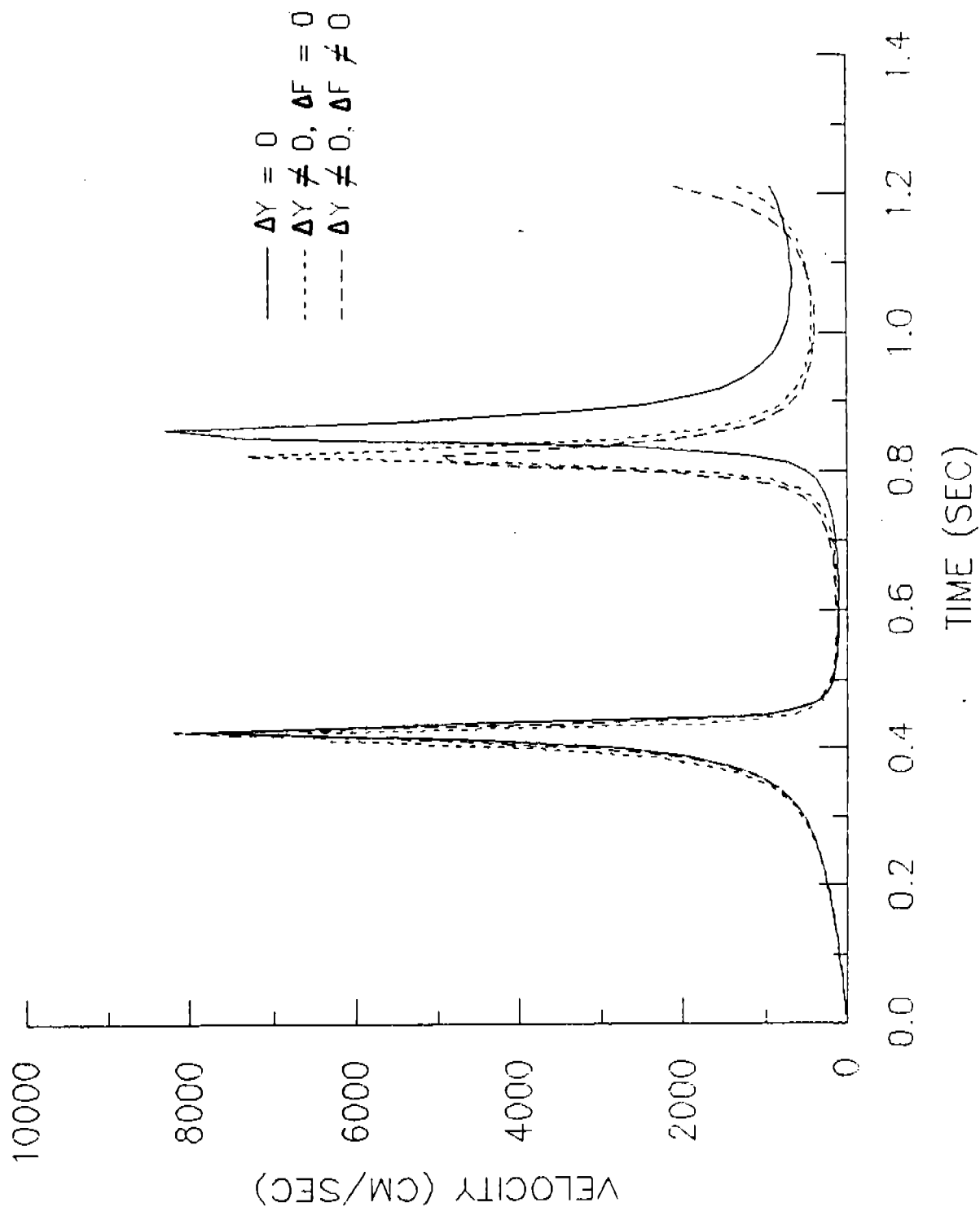


FIGURE 10

HEIGHT ABOVE EXPLOSION

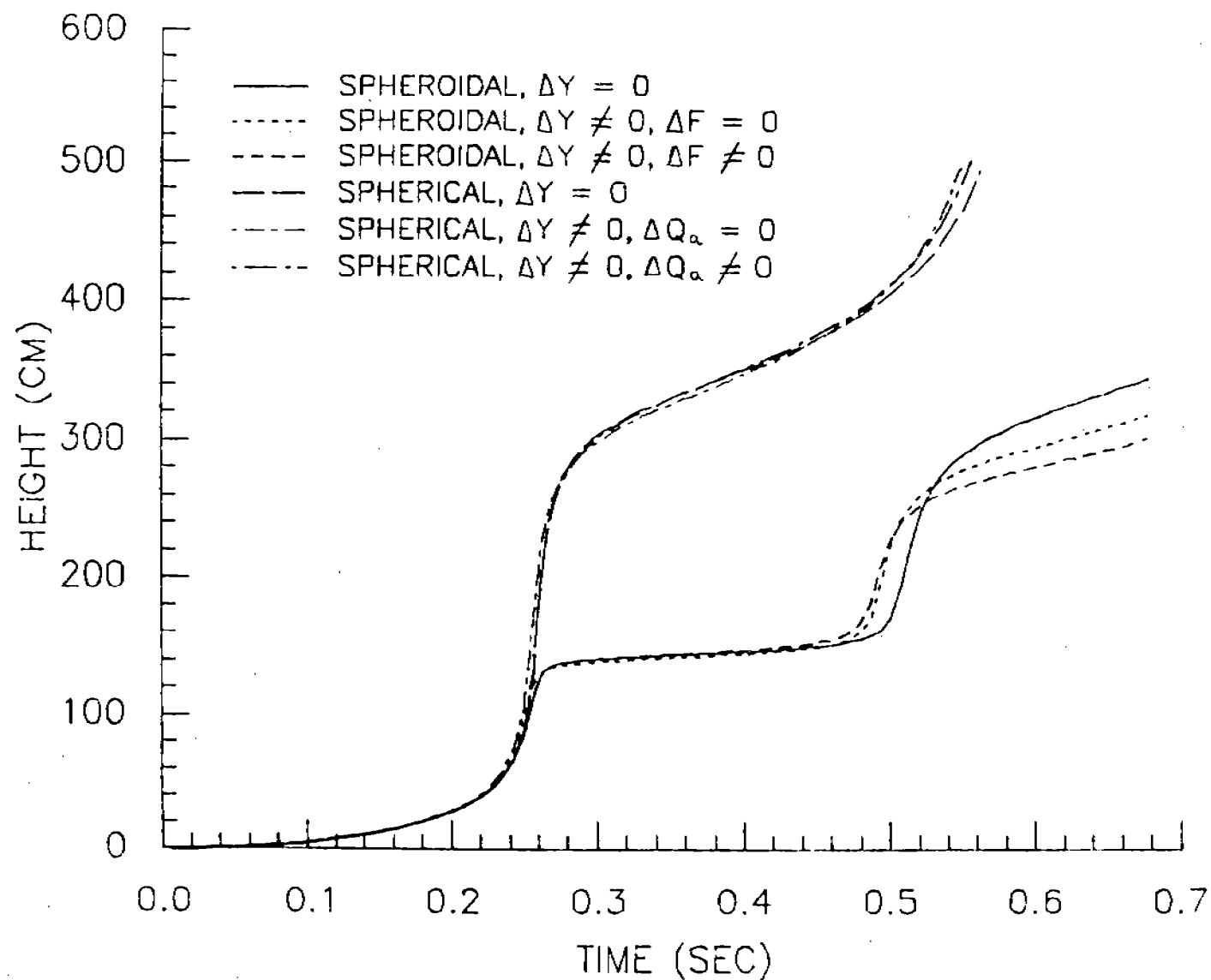


FIGURE 11

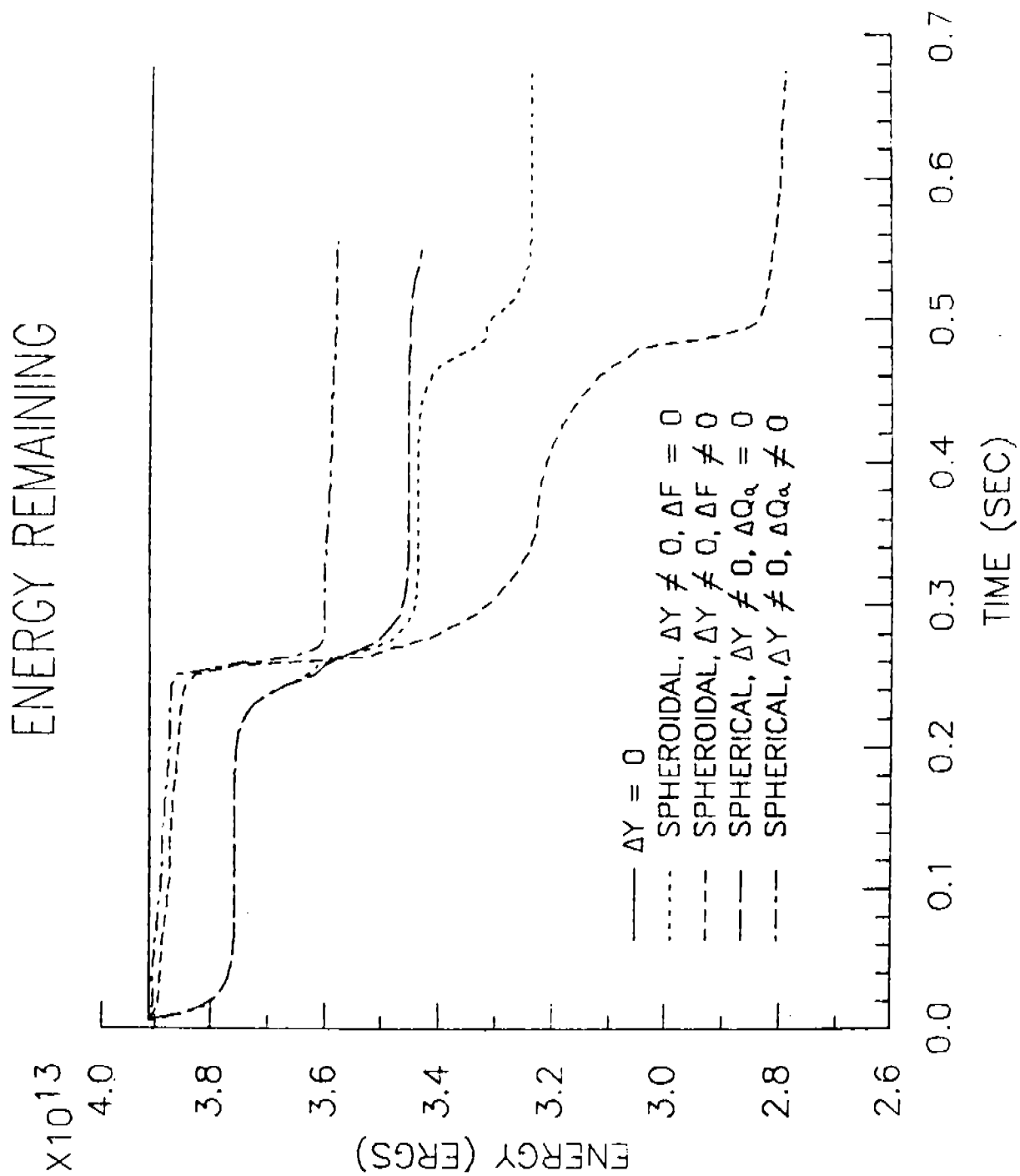


FIGURE 12

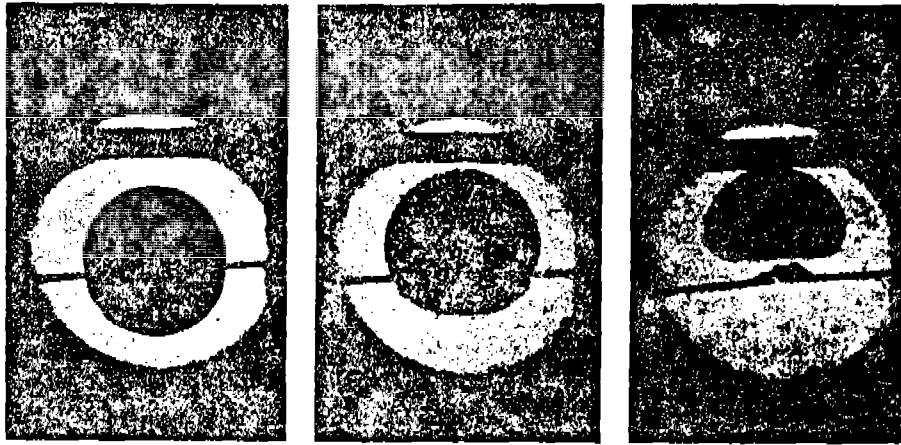


Fig. 4

Fig. 5

Fig. 6

Figs. 4-6. Photographs of bubbles in oil under vacuum, showing how the bubble becomes distorted as its age increases. Fig. 4. Spherical bubble, age 25 msec. Fig. 5. Bubble with pronounced flattening on underside, age 50 msec. Fig. 6. Mushroom-shaped bubble, age 80 msec.

(facing p. 346)

Evolution in time of a bubble in oil

(photograph from Taylor (1943))

FIGURE 13

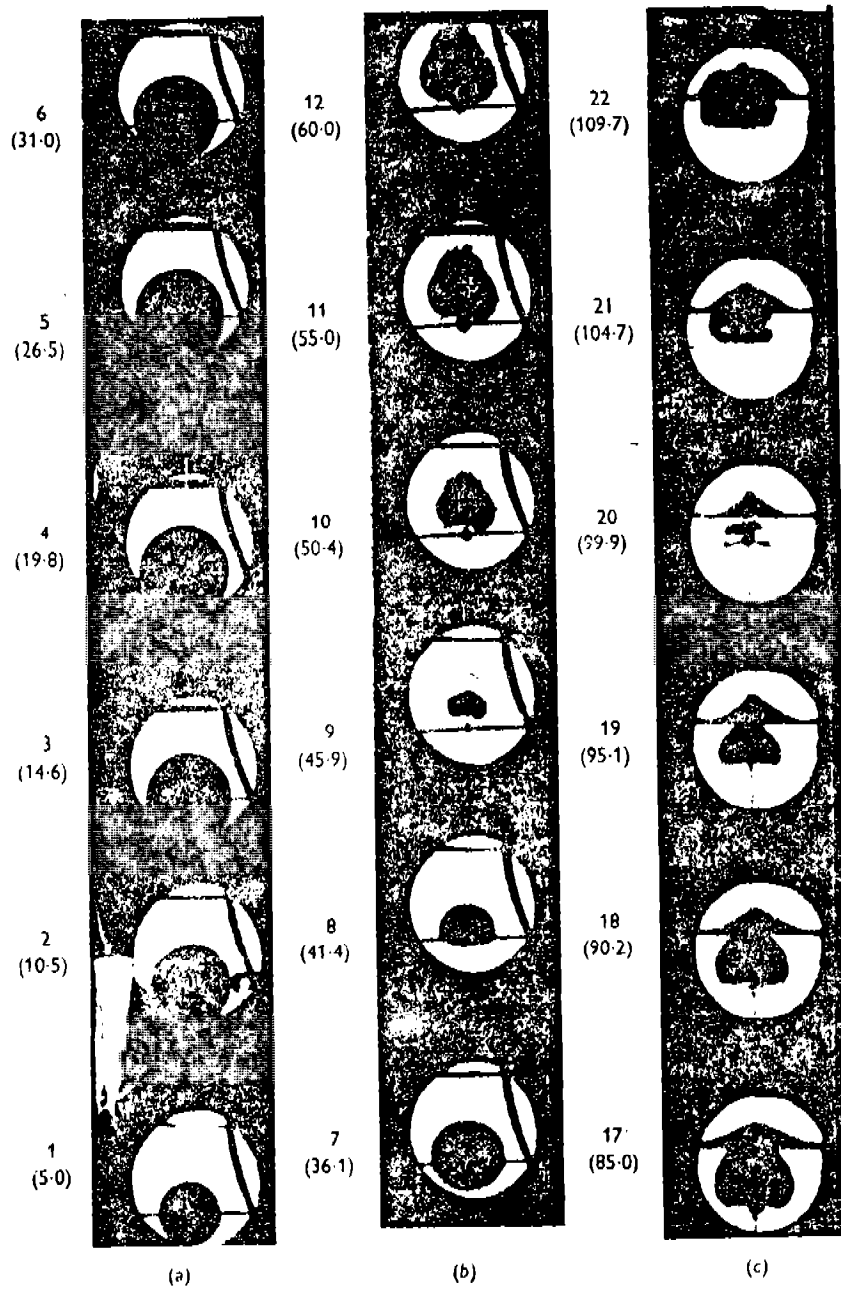


Fig. 8. Photographs of bubble in oil taken with a revolving drum camera. Upper number: identification number of photograph. Lower number (in parenthesis): age of bubble (in msec.).

Evolution in time of a bubble in oil
(photograph from Taylor (1943))

FIGURE 14

SEMI-AXES OF BUBBLE

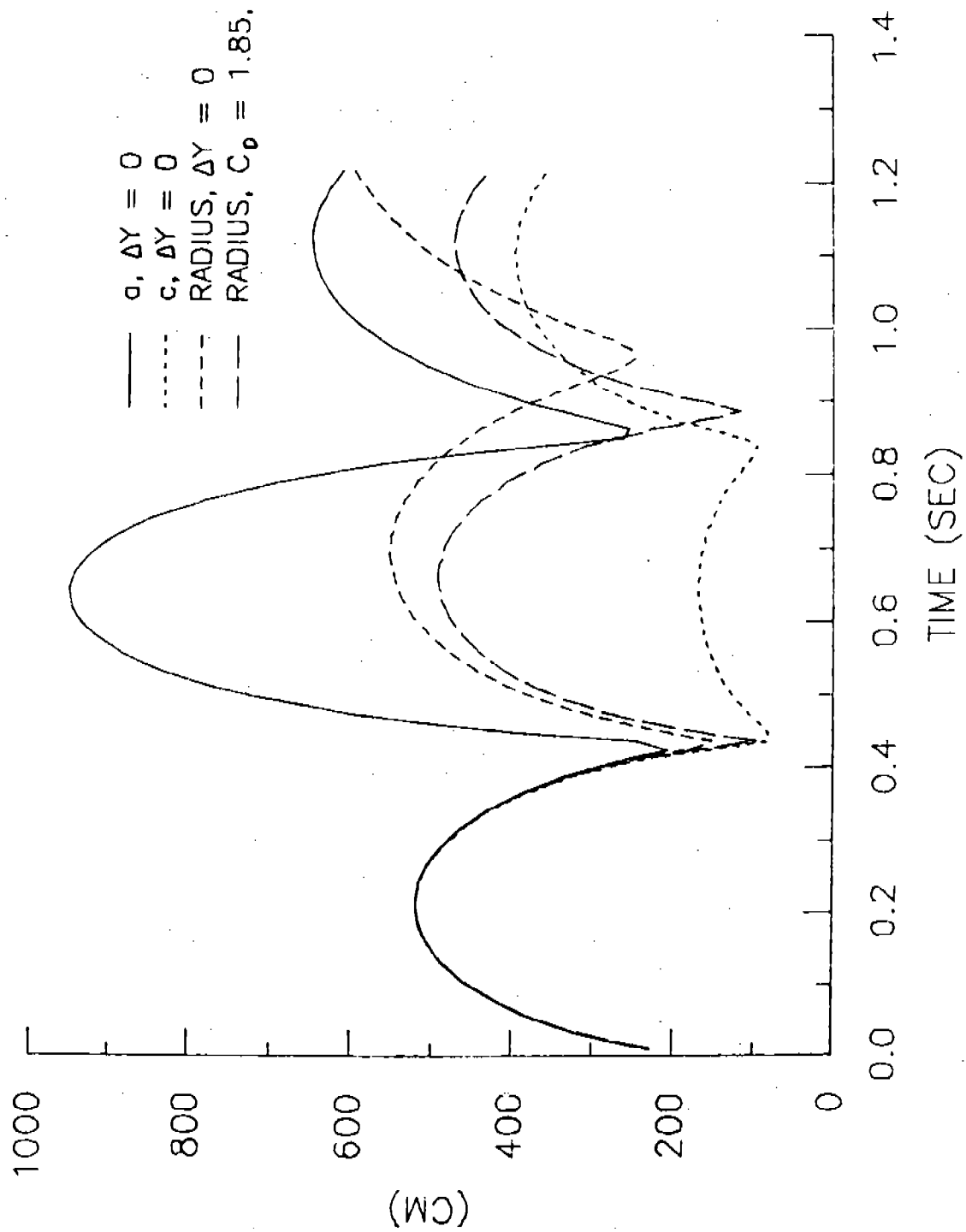


FIGURE 15

SEMI-AXES OF BUBBLE

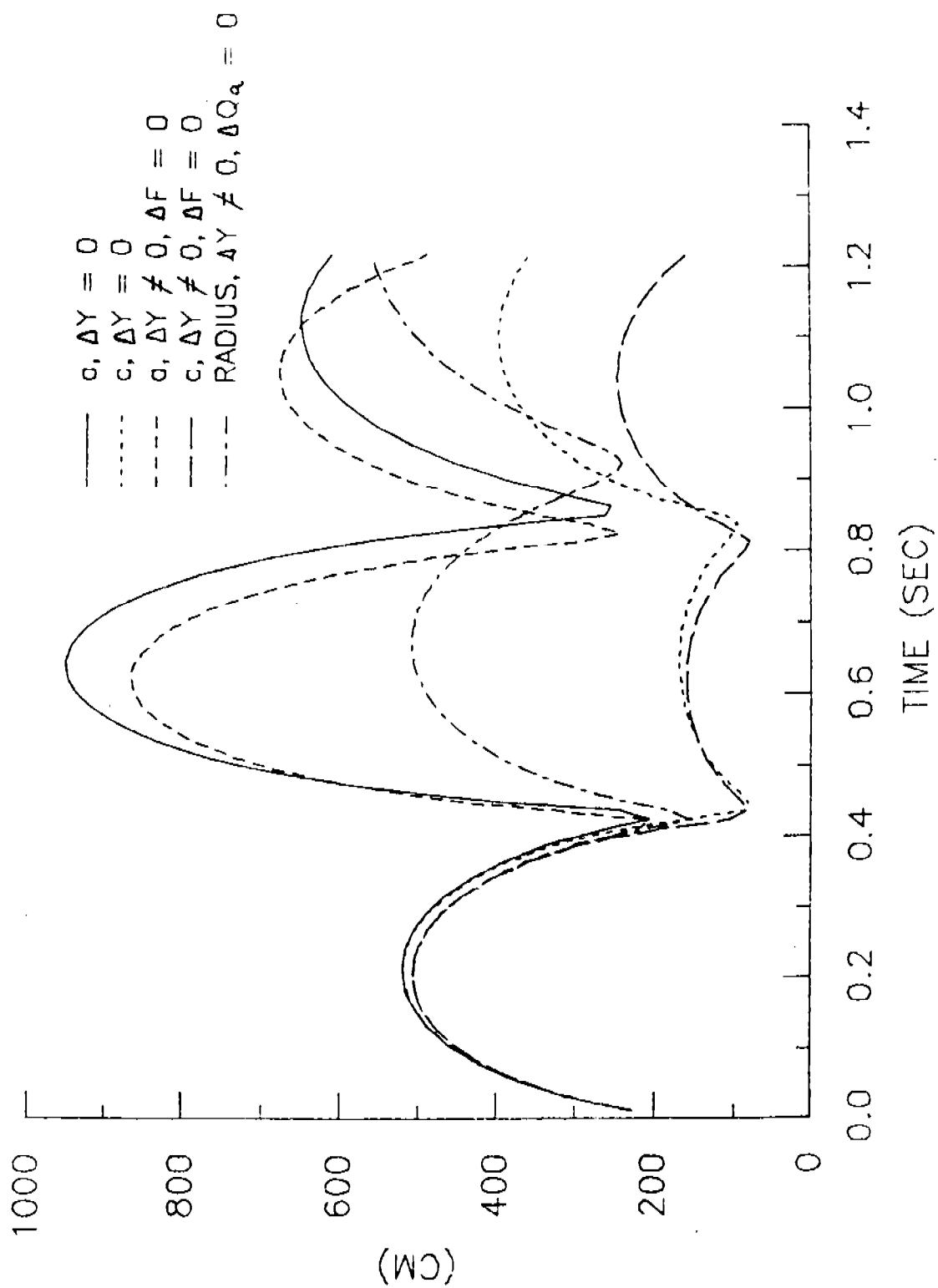


FIGURE 16

SEMI-AXES OF BUBBLE

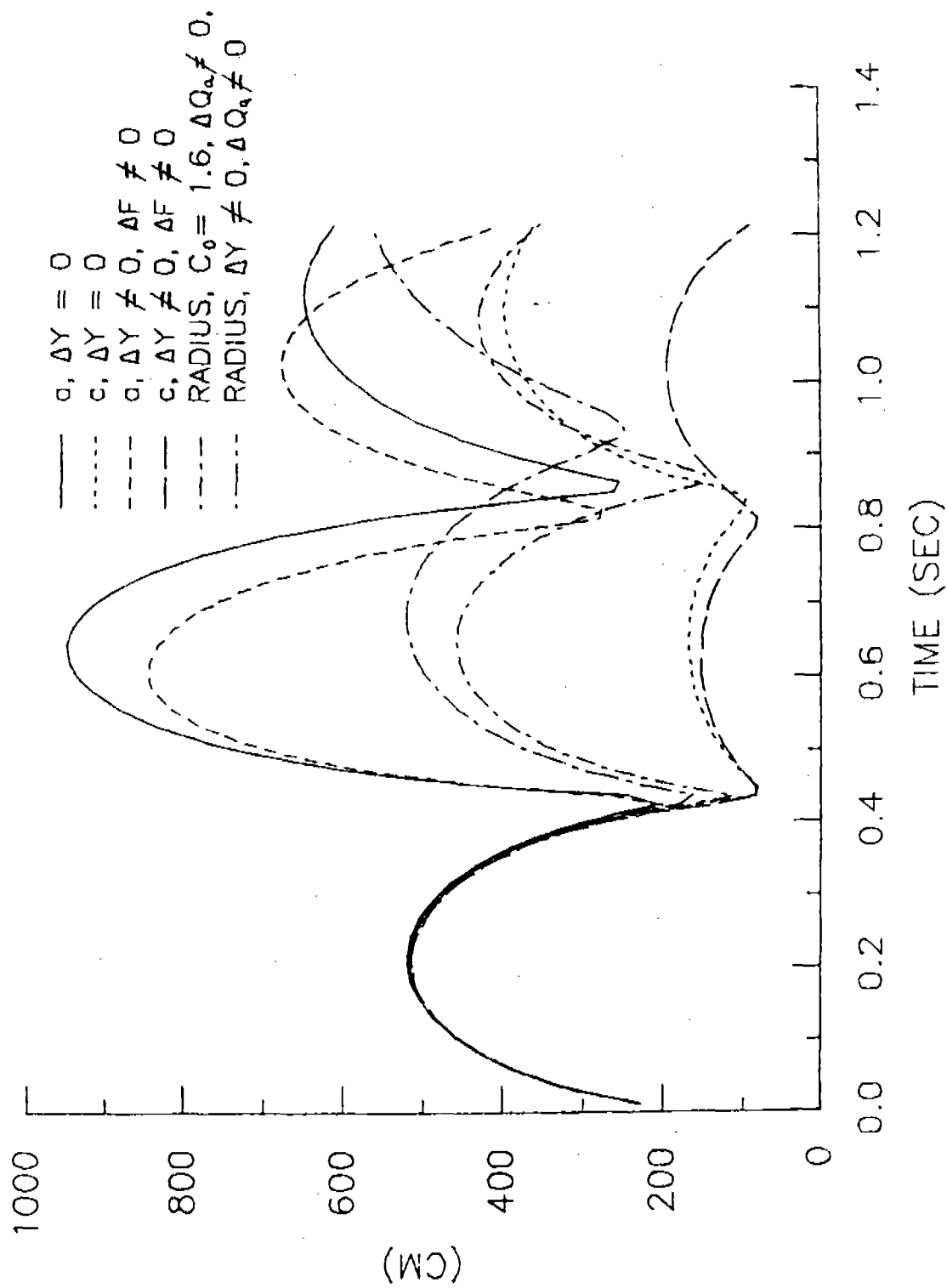


FIGURE 17

SEMI-AXES OF BUBBLE

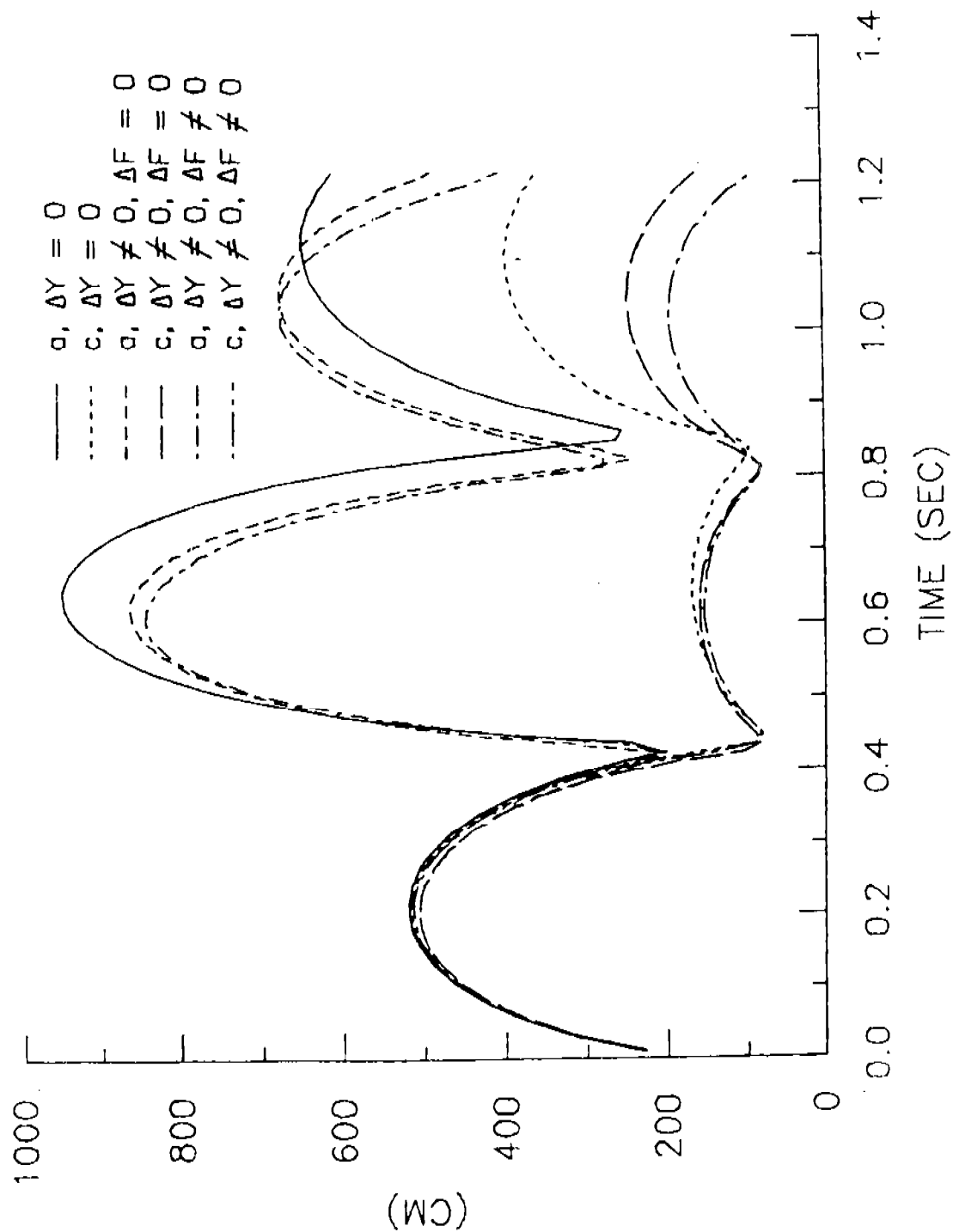


FIGURE 18

VELOCITY OF BUBBLE

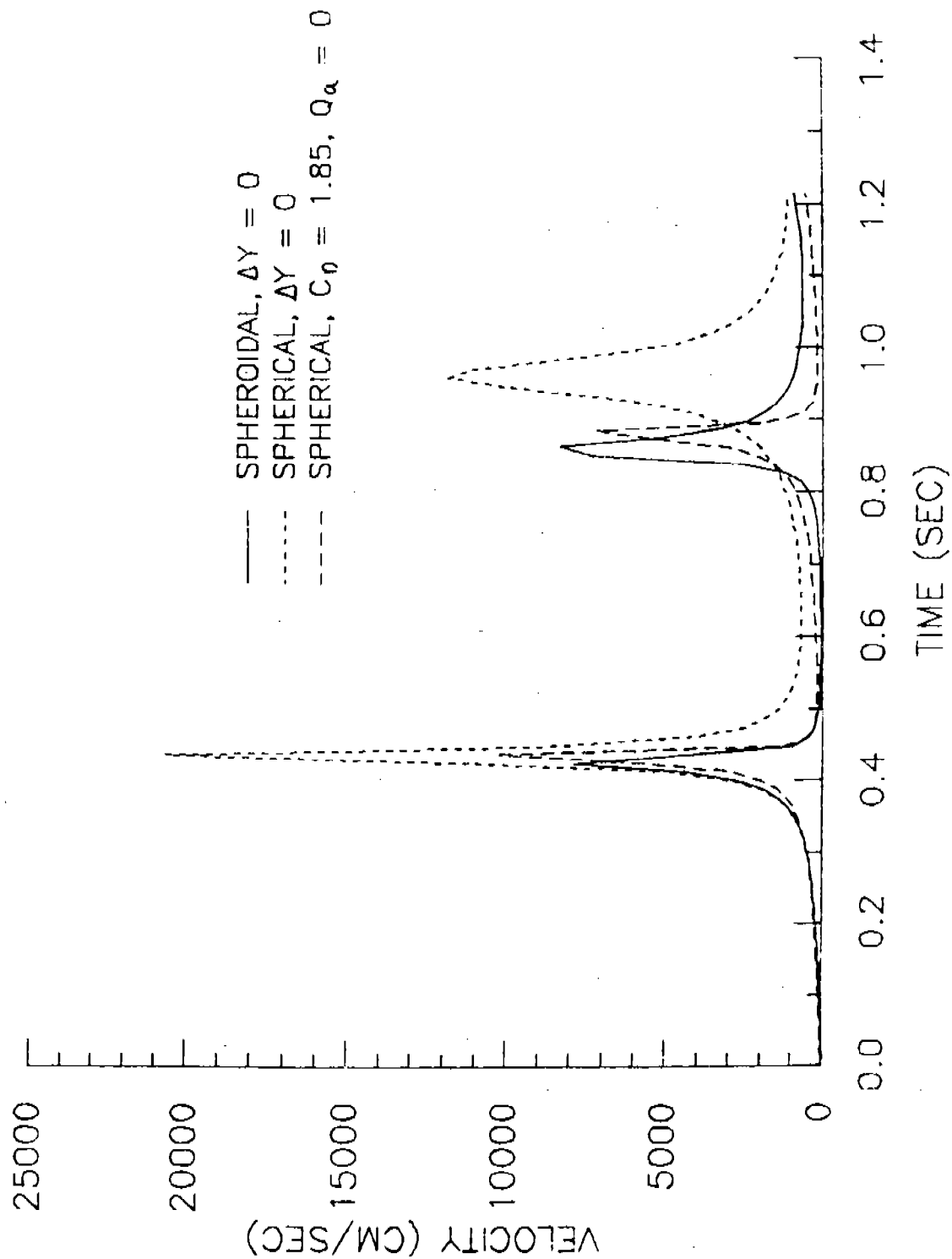


FIGURE 19

VELOCITY OF BUBBLE

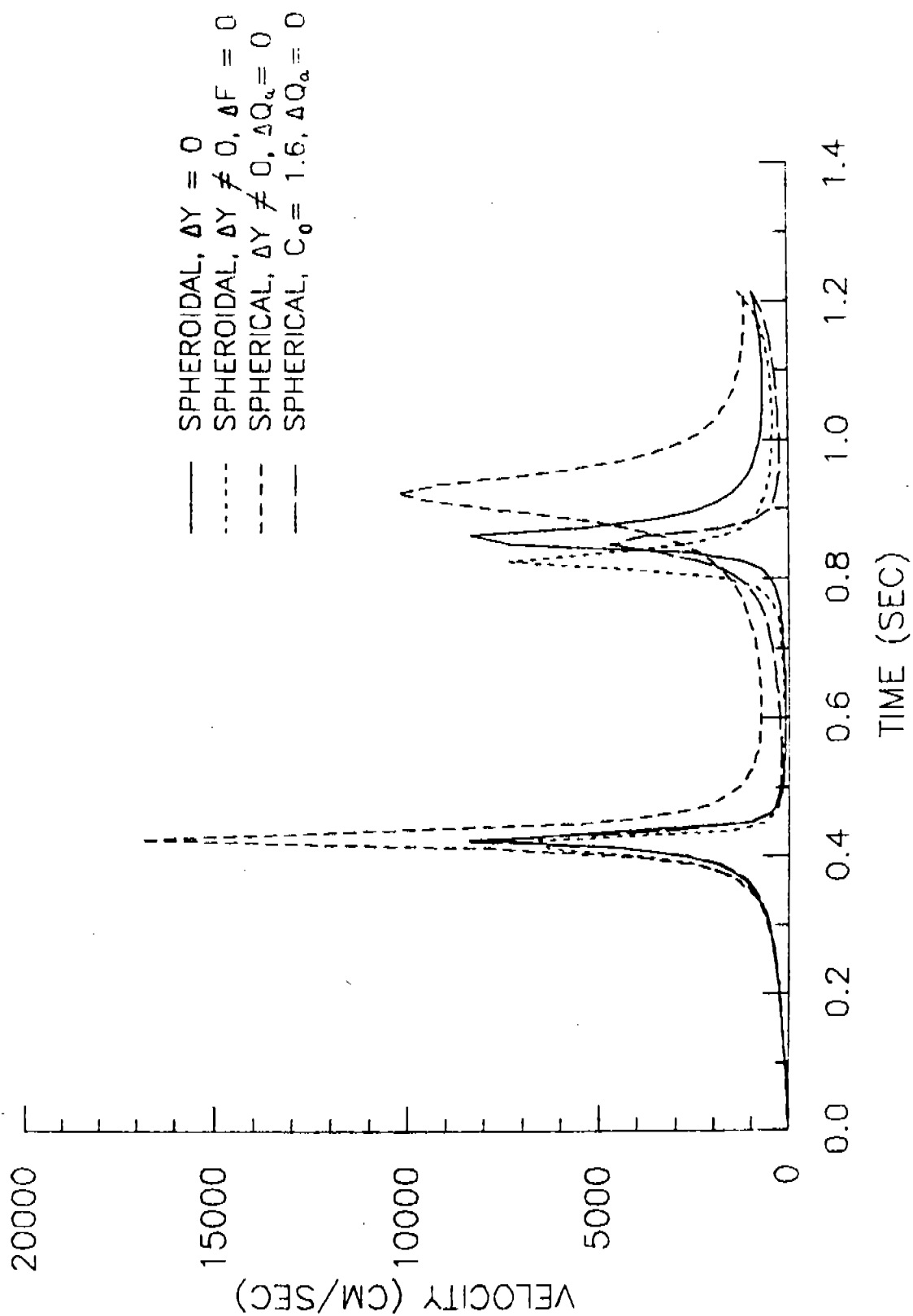


FIGURE 20

VELOCITY OF BUBBLE

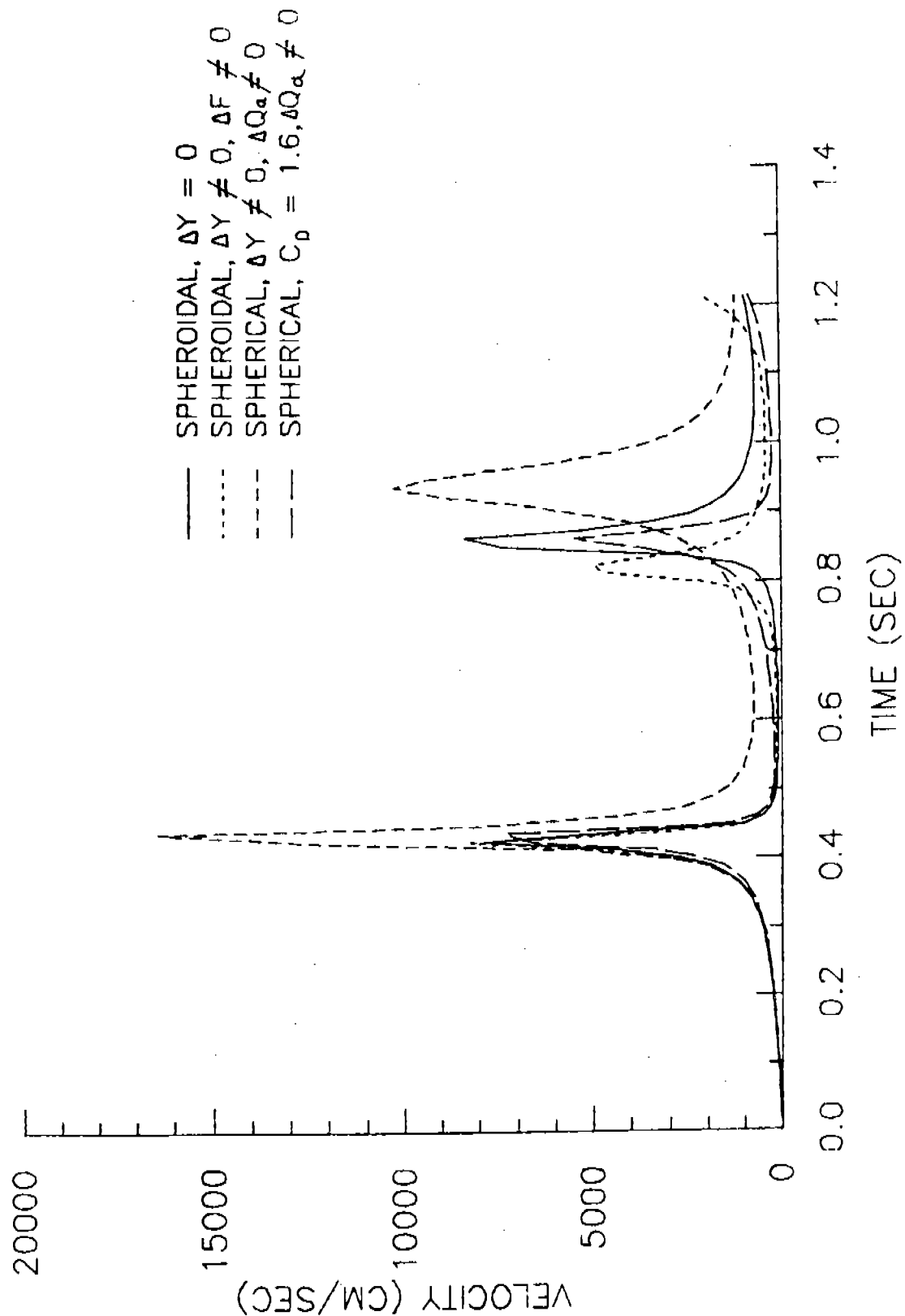


FIGURE 21

VELOCITY OF SPHEROIDAL BUBBLE

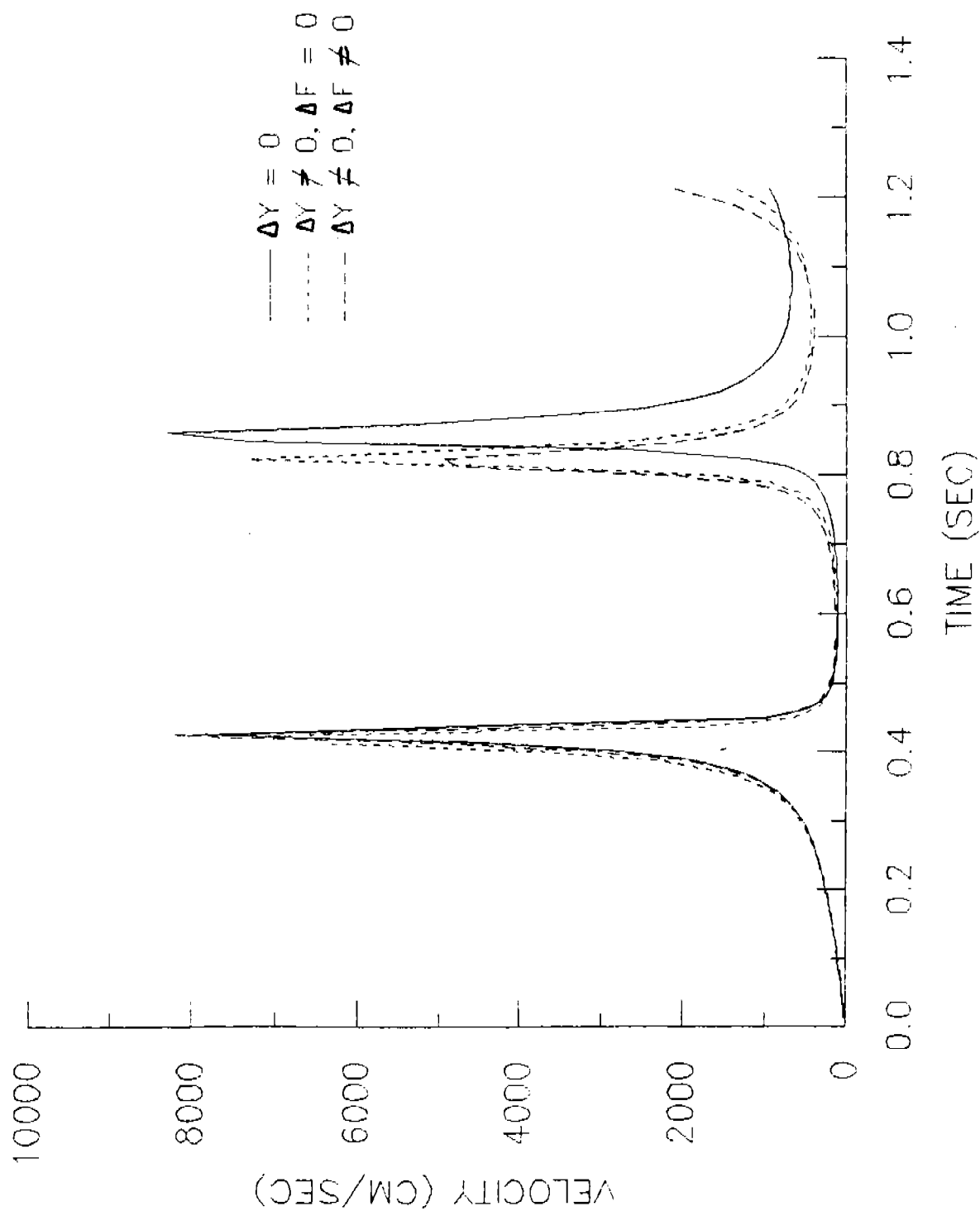


FIGURE 22

HEIGHT ABOVE EXPLOSION

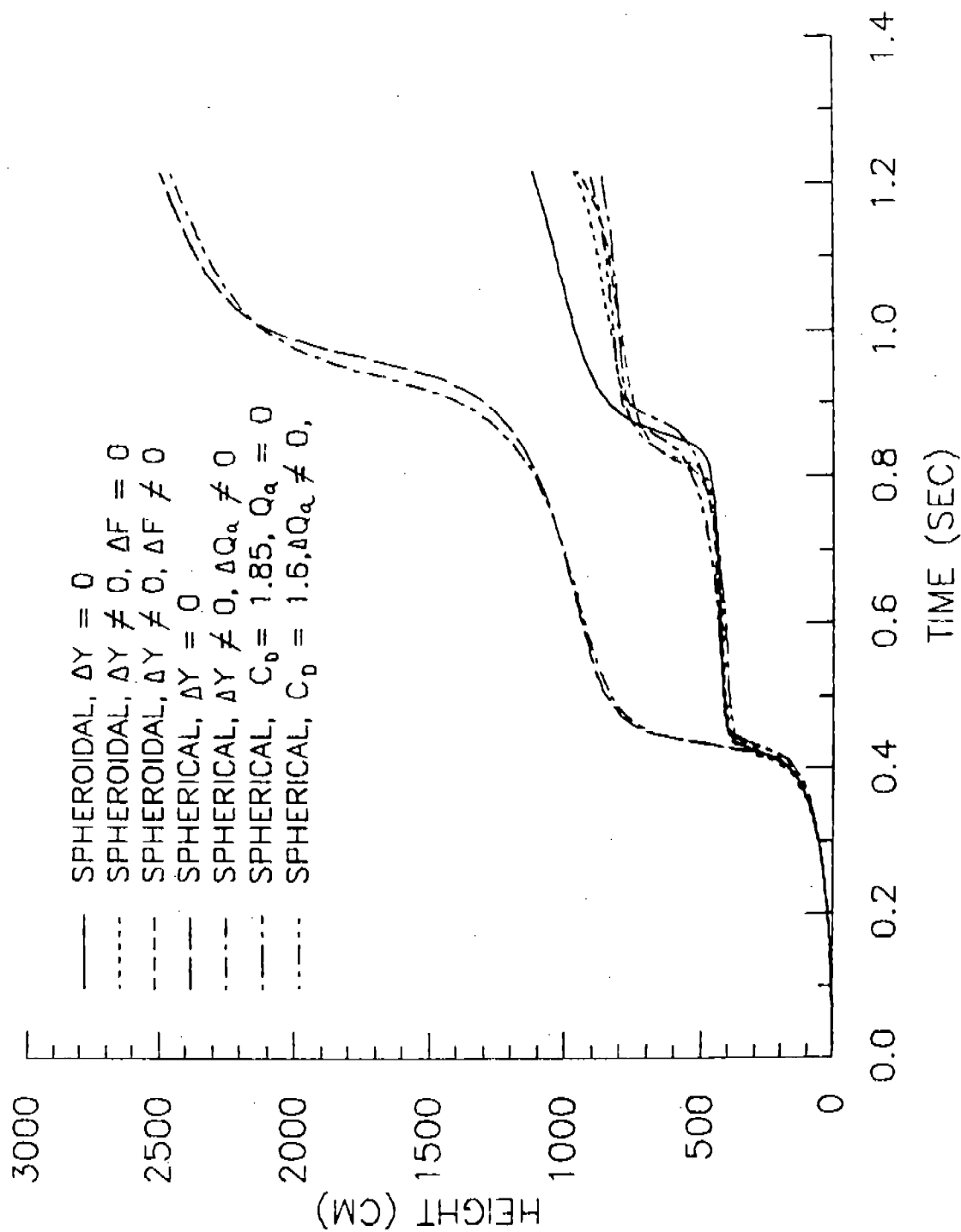
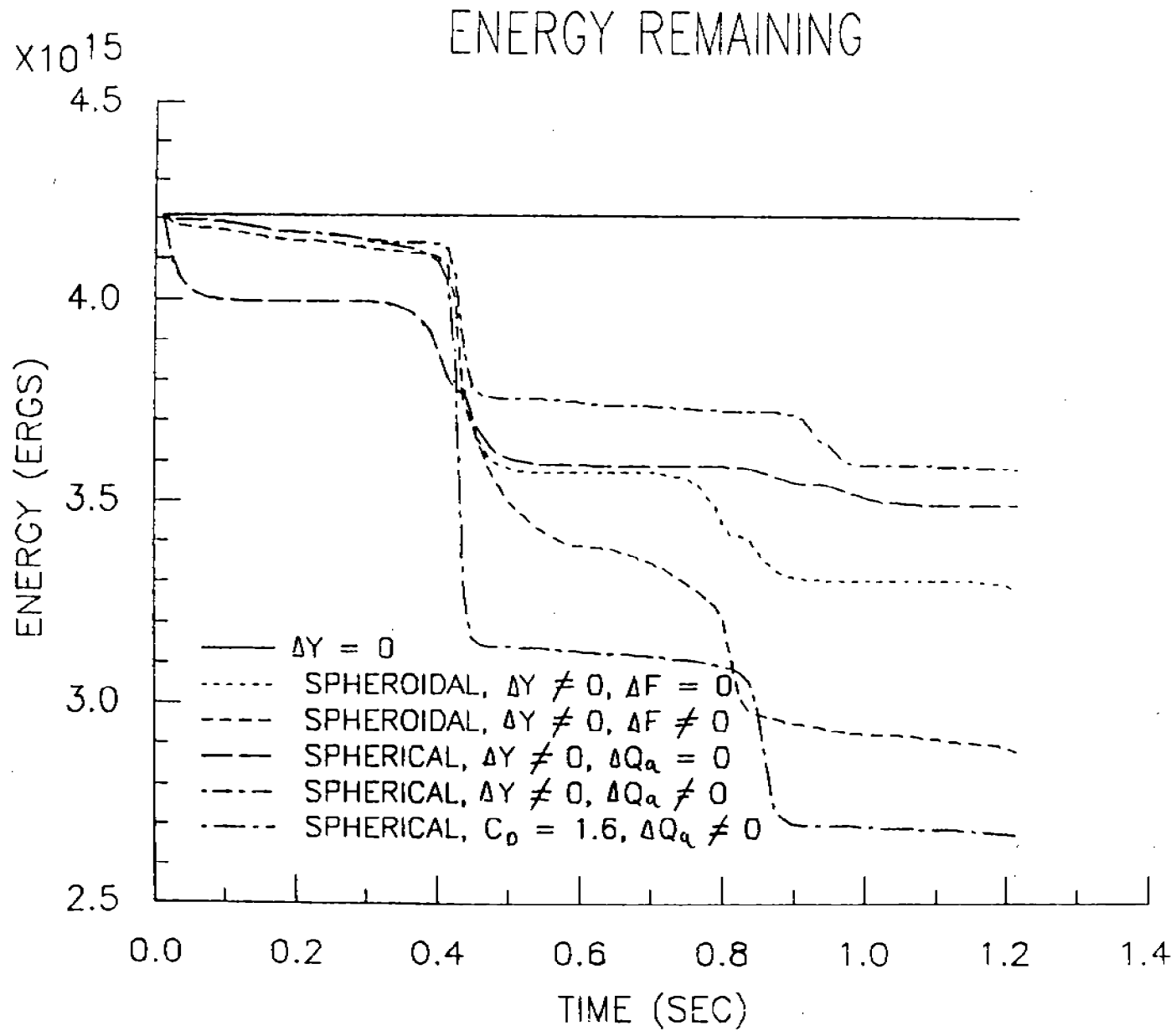


FIGURE 23



STEFAN'S PROBLEM IN A FINITE DOMAIN WITH CONSTANT BOUNDARY
AND INITIAL CONDITIONS

Shunsuke Takagi

U.S. Army Cold Regions Research and Engineering Laboratory
Hanover, N.H. 03755

ABSTRACT

Stefan's problem in a finite domain is solved under constant boundary and initial conditions. The solution is initially that of the semi-infinite domain, transits through infinitely many intermediate stage solutions, and finally arrives at a stationary stage solution. An intermediate stage solution emerges when a corresponding lead time becomes effective. Because exponential singularities exist at the terminal of the finite domain as functions of time, lead times are introduced.

The breakthrough is achieved by two innovations. The first is the use of the integral type solution of the one-dimensional heat conduction equation in place of the well-used serial type solution. By use of a finite domain solution determined under unrestricted initial and boundary conditions, the integral type solution allows us to find the temperature of the old phase in a two-phase Stefan problem. The second innovation is the inverse-Laplace-integral type expression of $i^k \text{erfc}(x/\sqrt{4\kappa t})$ that is valid for any integer k , negative, zero, or positive. This formula is used to expand the interfacial temperature of the old phase into a series of \sqrt{t} , and to sum up the series of the form,

$$\sum_{n=0}^{\infty} i^k \text{erfc} \frac{2n\ell \pm x}{\sqrt{4\kappa t}} .$$

This summation is employed to evaluate the final steady temperature.

NUMERICAL ABERRATIONS IN A STEFAN PROBLEM FROM DETONATION THEORY**

G.S.S. Ludford & A.A. Oyediran
Department of Theoretical & Applied Mechanics
Cornell University, Ithaca NY 14853 USA

ABSTRACT. The velocity of the moving boundary in a Stefan problem from detonation theory is examined analytically. Motivation comes from computations in which the velocity profile exhibited cusps and terminations (i.e. inability of the computer to find a velocity). Here we show that a solution exists with continuous velocity and acceleration at all times but that, under certain circumstances, another (singular) solution may bifurcate off. If a similar phenomenon occurs in the finite-difference schemes used, the analysis suggests that the aberrations (cusps and terminations) are due to numerical inaccuracies. The next step, apparently a difficult one, is to modify the schemes so as to follow the non-singular solutions.

I. INTRODUCTION. A certain model of detonation waves leads to the governing equations

$$F_t - [K(t) + F]F_x = F_{xx}, \quad F(-\infty, t) = 0, \quad F(+\infty, t) = F_\infty, \quad (1)$$

$$F(\mp 0, t) = F_*, \quad F_x(-0, t) - F_x(+0, t) = 1, \quad (2)$$

where F_*, F_∞ are given positive constants. Here x is measured from the moving boundary (the flame front), whose velocity is $K(t)$, and the solution must satisfy the given initial conditions

$$F(x, 0) = F_0(x) \text{ with } F_0(-\infty) = 0 \text{ and } F_0(+\infty) = F_\infty. \quad (3)$$

The problem (1-3) is overdetermined when the velocity of the boundary is prescribed; indeed, the object is to find $K(t)$ along with the solution $F(x, t)$. Numerically this is achieved by advancing the solution of the truncated system (1, 2a) at each time step, with K as a parameter whose value is determined by the remaining equation (2b).

There is a steady solution, expressible in terms of exponentials, whenever

$$2/F_\infty < F_* < F_\infty + 2/F_\infty, \quad (4)$$

but it is (linearly) unstable if

$$F_\infty/2 + 1/F_\infty + \sqrt{F_\infty^2/4 + 1/F_\infty^2} < F_* < F_\infty + 2/F_\infty. \quad (5)$$

Early computations [1] of unstable solutions give the curves in figure 1: eventually there was breakdown, i.e. the computer was unable to find a value of K , although in some cases this was preceded by the formation of cusps.

**Supported by the U.S. Army Research Office.

Later computations [2], based on more accurate numerical schemes always led directly to breakdown.

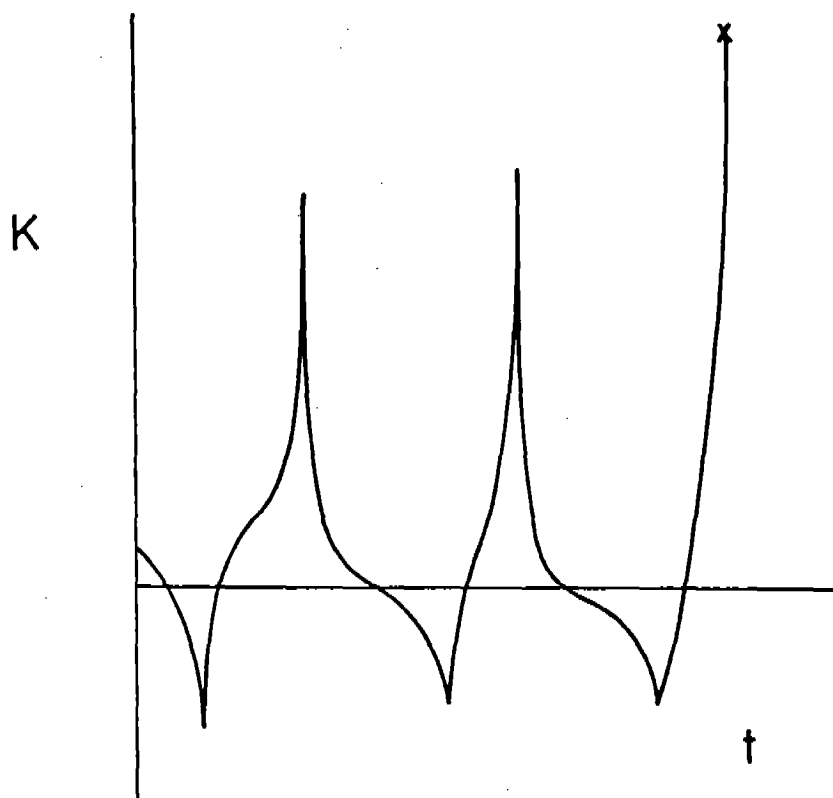


Figure 1. Computational aberrations for galloping detonation: cusps and termination.

The object of the present paper is to show that such cusps and breakdowns are numerical aberrations, i.e. at all stages of the evolution described by equations (1-3) the velocity $K(t)$ exists and the acceleration $K'(t)$ can be continuous. (An earlier paper [3] concluded that $K'(t)$ was discontinuous, but that is incorrect.) For that purpose we shall put the origin of t at the instant of interest, so that the initial conditions (3) are no longer arbitrary but must be the end result of the preceding evolution. This implies (Sec. 2) certain relations between the spacial derivatives on the two sides of the moving boundary, i.e. on the derivatives of F_0 at $x = \pm 0$. Such initial conditions are then shown (Sec. 3) to give the velocity the required properties

$$K(+0) = K(-0), K'(+0) = K'(-0). \quad (6)$$

Although there is always a solution with continuous velocity and acceleration, in certain circumstances there is a second (singular) solution. Section IV discusses the implications of this for numerical approximation of the evolution.

II. EVOLUTIONARY CONDITIONS. While initial values could be chosen so as to violate the boundary conditions (2), at any stage of the subsequent evolution these conditions must be satisfied. This implies

$$F_0(\mp 0) = F_*, \quad F'_0(-0) - F'_0(+0) = 1 \quad (7)$$

when $F_0(x)$ is the end result of an earlier evolution. Likewise, there are relations involving higher derivatives, in the generation of which the differential equation (1) is also involved. We shall need to consider derivatives up to the fifth order and, in doing so, the following consequences of the boundary conditions (2) will be used:

$$F^{\mp} = F_*, \quad F^{\mp}_t = F^{\mp}_{tt} = 0; \quad F^{\mp}_x - F^{\pm}_x = 1, \quad F^{\mp}_{xt} = F^{\pm}_{xt}, \quad F^{\mp}_{xtt} = F^{\pm}_{xtt} \quad (8)$$

where, for example,

$$F^{\mp}(t) = F(\mp 0, t). \quad (9)$$

The zeroth- and first-order relations (8a,d) lead to the requirement (7).

The second-order relation follows from the differential equation (1), which gives

$$-(K+F_*)F^{\mp}_x = F^{\mp}_{xx} \quad (10)$$

at $x = \mp 0$. Thus

$$F^{\mp}_{xx}/F^{\mp}_x = F^{\pm}_{xx}/F^{\pm}_x = V \text{ (say)}, \quad (11)$$

and the corresponding velocity is

$$K = -(F_* + V). \quad (12)$$

The third-order relation comes from the x -derivative of the differential equation, which gives

$$F^{\mp}_{xt} - (K+F_*)F^{\mp}_{xx} - F^{\mp 2}_x = F^{\mp}_{xxx} \quad (13)$$

at $x = \mp 0$. Thus,

$$[F_{xxx}] = V^2 - [F^2_x], \quad (14)$$

where the jump notation

$$[Q] = Q^- - Q^+ \quad (15)$$

(for any quantity Q) has been introduced.

To obtain the fourth-order relation, the second x - and first t -derivative of the differential equation must also be considered; these give, for $x = \mp 0$,

$$F_{xxt}^+ - (K+F_*)F_{xxx}^+ - 3F_x^+ F_{xx}^+ = F_{xxxx}^+, \quad (16)$$

$$-(K+F_*)F_{xt}^+ - K'F_x^+ = F_{xxt}^+. \quad (17)$$

Elimination of F_{xt}^+ , and F_{xxt}^+ from equations (13), (16) and (17) shows that

$$(F_{xxxx}^+ - 2VF_{xxx}^+ + 2VF_x^+ + V^3 F_x^+)/F_x^+ = (F_{xxxx}^+ - 2VF_{xxx}^+ + 2VF_x^+ + V^3 F_x^+)/F_x^+ = -A(\text{say}) \quad (18)$$

and the corresponding acceleration is

$$K' = A. \quad (19)$$

An alternative expression is

$$-A = [F_{xxxx}^+] - 2V[F_{xxx}^+] + 2V[F_x^+] + V^3. \quad (20)$$

Finally the fifth-order relation is obtained by also considering the third x-derivative and the second mixed derivative of the differential equation; these give, for $x = \bar{t}0$,

$$F_{xxxt}^+ - (K+F_*)F_{xxxx}^+ - 4F_x^+ F_{xxx}^+ - 3F_{xx}^+ = F_{xxxxx}^+, \quad (21)$$

$$F_{xtt}^+ - (K+F_*)F_{xxt}^+ - K'F_{xx}^+ - 2F_x^+ F_{xt}^+ = F_{xxxt}^+. \quad (22)$$

Elimination of F_{xt}^+ , F_{xxt}^+ , F_{xxxt}^+ from equations (13), (16), (21) and (22) leaves only the time derivatives F_{xtt}^+ , which may then be eliminated by subtraction. The result is

$$[F_{xxxxx}^+] = 3V[F_{xxxx}^+] - 6[F_x^+ F_{xxx}^+] + 7V^2[F_x^+] - 2[F_x^+] - 2V^4. \quad (23)$$

The requirement (8a) and the relations (8d), (11), (14), (18a), (23) between spacial derivatives on the two sides of the moving boundary provide restrictions on F_0 , if the initial data (3) are to be compatible with an earlier evolution. The velocity and acceleration attained just before the initial instant are given by the formulas (12) and (19) which, because they contain only space derivatives, may be written in terms of F_0 . Immediately after the initial instant the velocity and acceleration are given by the same formulas, so that whether or not they have discontinuities depends on whether or not the spacial derivatives do. Such infinitely rapid changes in derivatives at the boundary are effected by layers, so we turn now to their structure.

III. OUTER AND INNER EXPANSIONS. Let $t = 0$ be the instant at which a discontinuity is supposed to occur. We shall show that in fact

$$K_0 = K(+0), \quad K_1 = K'(+0) \quad (24)$$

exist and can be equal to

$$K(-0) = -(F_* + V), \quad K'(-0) = A, \quad (25)$$

where

$$V = C_-/G_- = C_+/G_+, \quad (26)$$

$$-A = (F_- - 2VT_- + 2VG_-^2 + V^3G_-)/G_- = (F_+ - 2VT_+ + 2VG_+^2 + V^3G_+)/G_+; \quad (27)$$

we may also write

$$-A = (F_- - F_+) - 2V(T_- - T_+) + 2V(G_- + G_+) + V^3. \quad (28)$$

The formulas (25) are rewrites of the results (12) and (19), while the expressions (26-28) come from the relations (11) and (18) and the expression (20) on setting

$$\begin{aligned} \bar{F}_x^+ &= F_0'(\bar{+}0) \equiv G_{\bar{+}}, \quad \bar{F}_{xx}^+ = F_0''(\bar{+}0) = C_{\bar{+}}, \quad \bar{F}_{xxx}^+ = F_0'''(\bar{+}0) \equiv T_{\bar{+}}, \\ \bar{F}_{xxxx}^+ &= F_0^{iv}(\bar{+}0) \equiv F_{\bar{+}}. \end{aligned} \quad (29)$$

We shall also need the rewrites

$$F_0(\bar{+}0) = F_*, \quad G_- - G_+ = 1, \quad (30)$$

$$T_- - T_+ = V^2 - (G_- + G_+), \quad (31)$$

$$F_-' - F_+' = 3V(F_- - F_+) - 6(G_-T_- - G_+T_+) + 7V^2(G_- + G_+) - 2(G_-^3 - G_+^3) - 2V^4 \quad (32)$$

of the relations (7), (14) and (23); we have set

$$\bar{F}_{xxxxx}^+ = F_0^v(\bar{+}0) \equiv F_{\bar{+}}'. \quad (33)$$

The strategy is to seek a solution with

$$K = K_0 + K_1\tau + \dots \quad \text{for } \tau > 0, \quad (34)$$

in which

$$F = F_0(x) + \tau F_1(x) + \tau^2 F_2(x) + \dots \quad (35)$$

away from the origin. The differential equation requires

$$F_1 = (K_0 + F_0)F_0' + F_0'', \quad (36)$$

$$2F_2 = (K_0 + F_0)^2 F_0'' + 2(K_0 + F_0)(F_0'^2 + F_0''') + 4F_0'F_0'' + F_0^{iv} + K_1 F_0'. \quad (37)$$

The relations (26b), (27b) (30-32) now ensure that

$$F_1(\bar{+}0) = F_2(\bar{+}0) = 0, \quad F_1'(-0) = F_1'(+0), \quad F_2'(-0) = F_2'(+0), \quad (38)$$

i.e. the boundary conditions (2) are satisfied to the order implied, provided the choices (25) are made for K_0 and K_1 .

The task of showing that K, K' exist and can be continuous at all times is therefore completed, but the proof (which essentially consists of two different ways of deriving the evolution relations) does not explain why discontinuities appear in the computations. For that we turn to a discussion of potential boundary layers, which uncovers an alternative (singular) solution in certain circumstances.

The boundary layers that must occur at the origin if there is a discontinuity are described by the similarity variable

$$\xi = x/t^{1/2}, \quad (39)$$

and we write

$$F = f_0(\xi) + t^{1/2}f_1(\xi) + tf_2(\xi) + t^{3/2}f_3(\xi) + t^2f_4(\xi) + t^{5/2}f_5(\xi) + \dots \quad (40)$$

Substitution in the differential equation (1) and the boundary conditions (2), plus matching with the expansion (35), then gives a series of differential problems for the coefficient functions $f_0, f_1, f_2, f_3, f_4, f_5$.

It was shown in [3] that, provided the initial data satisfy the relations (30), the first two functions in this expansion are

$$f_0 = F_*, \quad f_1 = G_{\mp} \xi \quad \text{for } \xi \leq 0. \quad (41)$$

The next term in the expansion satisfies

$$f_2'' + \frac{1}{2}\xi f_2' - f_2 = G_{\mp} K_*^* \quad \text{with } K_0^* = K_0 + F_*, \quad (42)$$

so that

$$f_2 = G_{\mp} K_*^* + A_{\mp}^{(2)} (\xi^2 + 2) + B_{\mp}^{(2)} e^{-\xi^2/8} D_{-3}(|\xi|/\sqrt{2}) \quad \text{for } \xi \leq 0. \quad (43)$$

Here D denotes the parabolic cylinder function and $A_{\mp}^{(2)}, B_{\mp}^{(2)}$ are integration constants, the latter representing the strengths of the boundary layers that effect instantaneous changes in the curvatures at the origin. Matching with the outer expansion (35) gives

$$A_{\mp}^{(2)} = C_{\mp}/2, \quad (44)$$

and then the boundary conditions (2) require

$$G_{\bar{+}}(K_0^*+V) + D_{-3}(0)B_{\bar{+}}^{(2)} = B_{-}^{(2)} + B_{+}^{(2)} = 0. \quad (45)$$

If $G_{-}+G_{+} \neq 0$, the (unique) solution of this homogeneous system is

$$K_0^* + V = B_{\bar{+}}^{(2)} = 0, \quad (46)$$

which confirms that K_0 exists and has the value (25a). Consideration of the problems for f_3 and f_4 then shows that K_1 has the value (25).

The cusps and terminations, found in the computations, correspond to

$$G_{-} + G_{+} = 0, \text{ i.e. } G_{\bar{+}} = \pm 1/2. \quad (47)$$

The solution of the homogeneous system (45) is then

$$B_{\bar{+}}^{(2)} = \bar{+}(K_0^* + V)/2 D_{-3}(0), \quad (48)$$

where K_0 (in K_0^*) is undetermined. To determine it we must go to the next term in the expansion (40), which satisfies

$$f_3'' + \frac{1}{2} \xi f_3' - \frac{3}{2} f_3 = -K_0^* f_2' - \frac{1}{4} \xi. \quad (49)$$

The general solution is

$$f_3 = K_0^* f_2' + \xi/4 + A_{\bar{+}}^{(3)} \xi(\xi^2+6) + B_{\bar{+}}^{(3)} e^{-\xi^2/8} D_{-4}(|\xi|/\sqrt{2}) \text{ for } \xi \gtrless 0, \quad (50)$$

where $A_{\bar{+}}^{(3)}, B_{\bar{+}}^{(3)}$ are integration constants, the latter enabling instantaneous changes in the third derivatives at the origin.

Matching with the outer expansion (35) gives

$$A_{\bar{+}}^{(3)} = T_{\bar{+}}/6, \quad (51)$$

and then the boundary conditions (2) require

$$-K_0^*(K_0^*+V)/\sqrt{\pi} + D_{-4}(0)B_{\bar{+}}^{(3)} = K_0^{*2} - T_{-} + T_{+} + D_{-4}'(0)(B_{-}^{(3)} + B_{+}^{(3)})/\sqrt{2} = 0. \quad (52)$$

Elimination of $B_{\bar{+}}^{(3)}$ from these three equations gives

$$K_0^{*2} + 3VK_0^* + 2(T_{-} - T_{+}) = 0 \quad (53)$$

where, according to the relation (31),

$$T_{-} - T_{+} = V^2. \quad (54)$$

It follows that either K_0 still has the value (25a) and

$$B_{\bar{+}}^{(3)} = 0, \quad (55)$$

or $K_0^* = -2V$ and $B_{\bar{+}}^{(2)}, B_{\bar{+}}^{(3)} \neq 0$.

We conclude that the solution bifurcates when $G_{\bar{+}} + G_{\bar{+}} = 0$: for one of the continuations, no boundary layers form at the origin and K is continuous; for the other, both the second and third derivatives are changed instantaneously at the moving boundary and there is a jump in velocity. The computational implications of this bifurcation will be discussed in Sec. IV; here we shall examine the first continuation further, with a view to confirming that the acceleration is then also continuous.

The next terms in the inner expansion are found to be

$$f_4 = V(3-4T_{\bar{+}})\xi^2/8 + V(1\pm V^2-4T_{\bar{+}})/4 \pm K_1/4 + F_{\bar{+}}(\xi^4+12\xi^2+12)/24 + B_{\bar{+}}^{(4)}e^{-\xi^2/8}D_{-5}(|\xi|/\sqrt{2}), \quad (56)$$

$$f_5 = -Vf_4' + (3V^2\pm 8T_{\bar{+}})\xi^3/24 + [6V^2-2V\pm 1-4(V^2\mp 3)T_{\bar{+}} \pm 2VK_1]\xi/8 + F_{\bar{+}}'\xi(\xi^4+20\xi^2+60)/120 + B_{\bar{+}}^{(5)}e^{-\xi^2/8}D_{-6}(|\xi|/\sqrt{2}) \quad (57)$$

after matching with the outer expansion. Applying the boundary conditions (2) to f_4 shows that

$$\pm(K_1-A)/4 + D_{-5}(0)B_{\bar{+}}^{(4)} = B_{\bar{-}}^{(4)} + B_{\bar{+}}^{(4)} = 0, \quad (58)$$

which should be compared with the system (45) for $G_{\bar{+}} = \pm 1/2$. As there the solution is not unique; we may write

$$B_{\bar{+}}^{(4)} = \mp(K_1-A)/4D_{-5}(0), \quad (59)$$

where K_1 is undetermined. To determine it we apply the boundary conditions to f_5 , and find

$$2V(K_1-A)/3\sqrt{\pi} + D_{-6}(0)B_{\bar{+}}^{(5)} = 3V(K_1-A)/\sqrt{2} - D_{-6}'(0)[B_{\bar{-}}^{(5)} + B_{\bar{+}}^{(5)}] = 0 \quad (60)$$

when equations (27), (28), (32) are used. Since $D_{-6}'(0)/D_{-6}(0) = -15\sqrt{\pi}/8\sqrt{2}$, this homogeneous system has the unique solution

$$K_1 = A, \quad B_{\bar{+}}^{(5)} = 0, \quad (61)$$

showing that the acceleration is indeed continuous.

REFERENCES

- [1] Stewart, D.S. & Ludford, G.S.S. Near Chapman-Jouget detonations. Transactions of the First Army Conference of Applied Mathematics and Computing, Washington D.C., 1983. ARO Report 84-1 (1984), pp. 801-811.
- [2] Oyediran, A.A. & Ludford, G.S.S. The Stefan problem of detonation theory. Transactions of the Second Army Conference on Applied Mathematics and Computing, Washington D.C., 1984. ARO Report 85-1 (1985), pp. 273-283.
- [3] Ludford, G.S.S. Saw-tooth evolution in a Stefan problem. To appear in Lectures in Applied Mathematics: Nonlinear Systems of PDE in Applied Mathematics.

THE ROLE OF MODELING IN AN INDUSTRIAL ENVIRONMENT

Vijay K. Stokes

General Electric Company
Corporate Research and Development

ABSTRACT

In many situations in industry -- especially those relating to the development of new concepts and machines -- the overall problem is not sufficiently well defined to warrant a "brute force" use of the field equations of mechanics. Such an approach can, in fact, be counterproductive, both from the point of view of the understanding achieved and the cost. It is more appropriate to first develop simple models to elucidate the physics of the problem. Once the basic mechanisms have been understood, more refined answers can be obtained by using the methodology of engineering science.

Two examples will be used to illustrate this point: A simple analytical model will be used to explain the apparently anomalous motion in a new orbital washer -- in which the motion of clothes is exactly opposite to what might be expected in such a machine. In the second example, simple beam theory will be used to develop an energy absorbing concept for thermoplastic automotive bumpers, which overcomes a major shortcoming in existing bumpers, namely their inability to absorb significant amounts of energy for impacts over the supports.

CONDENSATION ON FRACTALS SETS

J. S. Geronimo
School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332 USA

ABSTRACT. Among the many rich directions of study connected with fractal sets, there is the theory of moments of balanced measures on these sets. The balanced measure attaches equal weight to each of the mappings which generate the fractal set. This implies in many cases that the power moments can be calculated explicitly. One problem with balanced measures is that they are either absolutely or singularly continuous and consequently do not give rise to operators with eigenvalues. Here we report on a new class of sets and measures which we call condensed fractal sets and condensed measures which give rise to a point spectrum. We will also discuss some applications to specific physical problems.

I. INTRODUCTION. Let $R(z): \hat{C} \rightarrow \hat{C}$ denote a rational transformation of the extended complex plane $\hat{C} = C \cup \{\infty\}$ into itself of degree $N > 1$. Then $R(z) = P(z)/Q(z)$ where P and Q are nontrivial polynomials with no common factors. Set $R^{(n)}(z) = R \circ R^{(n-1)}(z)$ and $R^0(z) = z$. The Julia set J of $R(z)$ is equal to the closure of the set of all repulsive k cycles of R for all finite positive integers k .

Associated with the action of R on \hat{C} there is a unique measure μ , called the balanced measure for R , with the property that

$$\int_J f d\mu = \frac{1}{N} \int_J \sum_{i=1}^N f(R_i^{-1}(x)) d\mu \quad (1)$$

where $f \in L^1(J, \mu)$ and $R_i^{-1}(z)$, $i = 1, 2, \dots, N$, denotes a complete assignment of the branches of $R^{-1}(z)$. If $R(z)$ is a polynomial then the orthonormal polynomials $p_n(x)$ associated with μ i.e., $p_n(x) = k(n)x^n + \dots$, $k(n) > 0$ with

$$\int_J p_n(x) \overline{p_m(x)} d\mu = \delta_{n,m} \quad (2)$$

satisfy the relation [BGH1], [BM]

$$p_{nN}(x) = p_n(Tx) \quad (3)$$

When J is real these polynomials also satisfy the following three term recurrence formula

$$a(n+1)p_{n+1}(x) + b(n)p_n(x) + a(n)p_{n-1}(x) = xp_n(x) \quad (4)$$

with $p_0(x) = 1$ and $p_{-1}(x) = 0$. The above is just the discrete analog of the Schrödinger equation. Let A be the infinite dimensional Jacobi matrix associated with (4), i.e.,

$$A = \begin{pmatrix} b(0) & a(1) & & \\ a(1) & b(1) & a(2) & \\ & a(2) & b(2) & \ddots \\ & & \ddots & \ddots \end{pmatrix}$$

then a consequence of (3) is that A satisfies the following renormalization group equation [B], [BGM],

$$D(z-A)^{-1}D^* = (Tz-A)^{-1}. \quad (5)$$

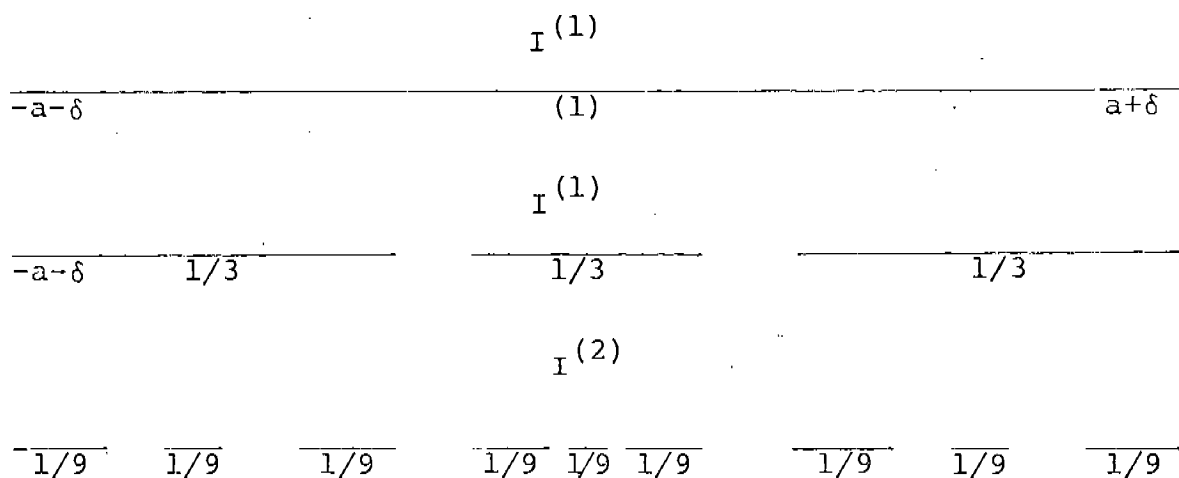
Here $D: \ell_2^+ \rightarrow \ell_2^+$ and $(D\psi)(n) = \psi(nN)$, $\psi \in \ell_2^+$.

II. CONDENSED JULIA SETS. Instead of diving into the general phenomena of condensation we will illustrate it with a simple example. Let us consider the Julia sets J and J_ϵ associated with the transformations $Rz = z^2 - \lambda$ and $R_\epsilon z = z^2 - \lambda + \epsilon/z$ respectively. Here $\lambda > 2$. We begin by recalling the construction of J . Let $I^{(0)} = [-a, a]$ with $a = \frac{1}{2} + \sqrt{\lambda + \frac{1}{4}}$, this being the largest fixed point of R . Defining $I^{(n)} = R^{-1}I^{(n-1)}$, it is readily shown that $I^{(n)} \subset I^{(n-1)}$ and that $J = \lim_{n \rightarrow \infty} I^{(n)}$. $I^{(n)}$ for $n = 0, 1$, and 2 are shown below.

$$\begin{array}{c} I^{(0)} \\ \hline -a \qquad \qquad \qquad (1) \qquad \qquad \qquad a \\ \\ I^{(1)} \\ \hline -a \qquad \qquad -\sqrt{\lambda-a} \qquad \qquad \sqrt{\lambda-a} \qquad \qquad a \\ \qquad (1/2) \qquad \qquad \qquad \qquad \qquad \qquad (1/2) \\ \\ I^{(2)} \\ \hline -a \qquad \qquad -\sqrt{\lambda-a} \qquad \qquad \sqrt{\lambda-a} \qquad \qquad a \\ \qquad (1/4) \qquad \qquad (1/4) \qquad \qquad (1/4) \qquad \qquad (1/4) \end{array}$$

A sequence of measures $\{\mu^{(n)}\}_{n=0}^{\infty}$ which converge weakly to μ can be constructed. One begins by letting $\mu^{(0)}$ have total mass 1 on $I^{(0)}$. Now following (1) $\mu^{(1)}$ will have 1/2 its mass on each of the two intervals of $I^{(1)}$, $\mu^{(2)}$ will have 1/4 its mass on each of the four intervals of $I^{(2)}$ etc. etc.

Let us construct J_{ϵ} when $0 < \epsilon \ll 1$. Let $I^{(0)} = [-a-\delta, a+\delta]$, where $a+\delta$ is the largest positive fixed point of $R_{\epsilon}(z)$, so $\delta \rightarrow 0$ as $\epsilon \rightarrow 0$. Let $I^{(n)} = R_{\epsilon}^{-1} I^{(n-1)}$ then again $J_{\epsilon} = \lim_{n \rightarrow \infty} I^{(n)}$. $I^{(n)}$ for $n = 0, 1, 2$ are shown below



Since R_{ϵ} has three inverse branches we see that the original interval $I^{(0)}$ will be split into three disjoint intervals for ϵ small enough, and each of these intervals will in turn split into three other disjoint intervals.

Again one can construct a sequence of measures $\{\mu^{(n)}\}$ which converge weakly to μ_{ϵ} . Here $\mu_{\epsilon}^{(0)}$ has total mass 1 on $I^{(0)}$ while $\mu_{\epsilon}^{(n)}$ places a total mass of $1/3^n$ on each of the 3^n intervals of $I^{(n)}$. In the limit as $\epsilon \rightarrow 0$ we see that

$$J_0 = J \cup \{\lim_{n \rightarrow \infty} R^{-n}(0)\}. \quad (6)$$

That is J_0 is the Julia set associated with the transformation $R(z) = z^2 - \lambda$ plus zero and all its preimages. Furthermore the measure associated with J_0 becomes

$$d\mu_0 = \sum_{n=0}^{\infty} \sum_{m=1}^{2^n} \frac{1}{3^{n+1}} \delta(x - x_m^{(n)}) dx \quad (7)$$

where $\{x_m^{(n)}\}_{m=1}^{2^n}$ denotes the preimages of order n under $R(z)$ of zero, and $\delta(x)$ is the Dirac delta function. We note the analog of (1) for μ_0 is

$$\int_{J_0} f d\mu_0 = \frac{1}{3} \int_{J_0} \sum_{i=1}^2 f(R_i^{-1}x) d\mu_0 + \frac{1}{3} f(0) \quad (8)$$

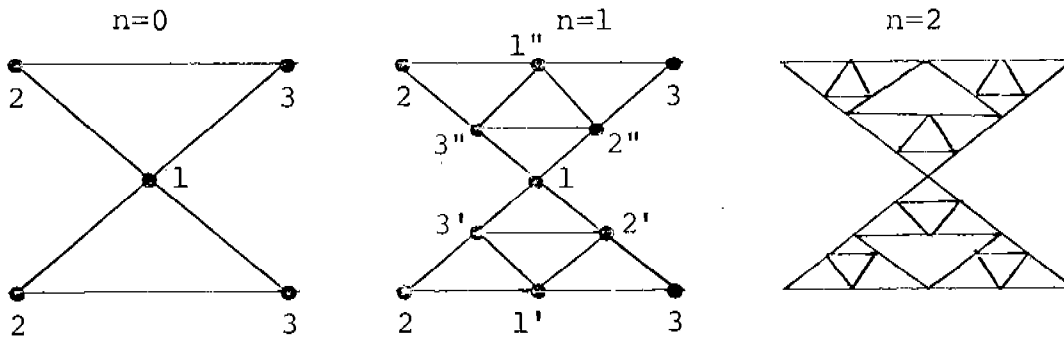
$f \in L^1(J, \mu_0)$.

Generalizations of the above discussion to more general rational functions can be found in [BGH2] and to more general mappings of compact spaces into themselves in [BD].

III. APPLICATIONS.

a) Schrödinger equation on a Sierpinski lattice [DABK], [R].

We consider the sequence of fractal lattices shown below:



The boundary conditions used identify the corners of the two triangles on the largest scale. Stopping at any finite n , we consider the tight-binding eigenvalue equation in ℓ_2

$$H|\psi\rangle = E|\psi\rangle \quad (9)$$

where H is given by

$$H = - \sum_{\substack{\text{nearest} \\ \text{neighbors}}} \{ |i\rangle \langle j| H |j\rangle \langle i| \}$$

It can be shown [BGH2] that the spectrum of H is $C \cup \{-4, 2\}$, where C is the condensed Julia set for $z \mapsto -z(z+3) +$

$\epsilon(z^2-1)^{-1}$. Let σ_1 and σ_2 be the condensed measures for $z \mapsto -z(z+3)(z+1)/(z+1)$ and $z \mapsto -z(z+3)(z-1)/(z-1)$ respectively; then the density of states ρ associated with H can be written as

$$\rho = \frac{1}{3} \sigma_1 + \frac{1}{3} \sigma_2 + \frac{1}{3} \chi_{-2}.$$

Here χ_{-2} is the characteristic function of the set $\{-2\}$.

b) An application to orthogonal polynomials. Let $\{p_n\}$ be the polynomials given by (2) and assume that J is real. We now construct the polynomials of the second kind

$$p_n^{(1)}(x) = \int_J \frac{p_{n+1}(x) - p_{n+1}(y)}{x - y} d\mu(y).$$

These polynomials obey the relation [BGH3]

$$p_{nN-1}^{(1)}(x) = \frac{R'(x)}{N} p_{n-1}^{(1)}(R(x)).$$

Furthermore the measure with respect to which these polynomials are orthogonal is a condensed measure and one has [BGM]

$$\begin{aligned} \int_C f d\mu^{(1)} &= \sum_{i=1}^N \int_C \frac{Nf(R_i^{-1}(x)) d\mu^{(1)}}{R'(R_i^{-1}(x))^2} \\ &+ \sum_{n=1}^N \Gamma_n f(x_n), \quad f \in L^1(C, \mu^{(1)}). \end{aligned}$$

Here C is the condensed Julia set associated with $\mu^{(1)}$, $\{x_n\}_{n=1}^N$ are the zeros of $R'(x)$, and $\{\Gamma_n\}_{n=1}^N$ are predetermined constants [BGM].

References

- [BD] M. F. Barnsley and S. Demko, Proc. Roy. Soc. London (accepted).
- [BGH1] M. F. Barnsley, J. S. Geronimo, and A. N. Harrington, Bull. AMS 7, 381 (1982).
- [BGH2] M. F. Barnsley, J. S. Geronimo, and A. N. Harrington, Trans. Amer. Math. Soc. 2, 537 (1985).
- [B] J. Bellissard, Lecture Notes in Math., Springer-Verlag, to appear.
- [BGM] D. Bessis, J. S. Geronimo, and P. Moussa, Commun. Math. Phys., submitted.
- [BM] D. Bessis and P. Moussa, Commun. Math. Phys. 88, 503 (1983).
- [DABK] E. Domany, S. Alexander, D. Bensimon, and L. P. Kadanoff, Phys. Rev. B 28, 3110 (1983).
- [R] R. Rammal, J. Physique (Paris), 45, 191 (1984).

NEWTON'S METHOD, JULIA SETS AND CHAOTIC DYNAMICS

Edward R. Vrscay
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

The Schröder or König iteration schemes for a given polynomial $g(z)$ represent generalizations of Newton's method. In both schemes, functions $S_m(z)$ are constructed so that the iteration sequence $z_{n+1} = S_m(z_n)$ converges locally to a root z^* of $g(z) = 0$ with prescribed order m . Pathological situations involving the iteration sequence $\{z_n\}$ do exist, however. If $z_0 \in J$, the Julia set of rational function $S_m(z)$, then the z_n behave chaotically and never converge to a root. For $z_0 \notin J$, it is also possible, however, that the z_n converge to attractive cycles other than the roots z^* . Associated with this latter behavior are regions in a "parameter space" which exhibit the morphology and dynamical patterns which are associated with the classical Mandelbrot sets of quadratic maps. These types of behavior are investigated with the aid of microcomputer plots.

1. Introduction

Consider the familiar and very important Newton iteration function, constructed to determine the roots of a polynomial $g(z)$:

$$N(z) = z - \frac{g(z)}{g'(z)}. \quad (1.1)$$

Clearly, $N(z)$ is a rational function. One is interested in the conditions for which the iteration sequence $z_{n+1} = N(z_n)$ converges to a root z^* of $g(z)$. We may ask several questions: What is the set $W(z^*)$ of all initial values $z_0 \in \mathbb{C}$ for which the sequence $\{z_n\}$ converges to the root z^* ? Is it possible that the z_n do not converge to any of the z_i^* ? For what $z_0 \in \mathbb{C}$ do these pathological situations occur and what is the nature of these nonconvergent sequences $\{z_n\}$? The classical theory of iteration of analytic functions, concerned with the behavior of sequences in the neighborhood of a fixed point, does not attack such global problems. Here, the Julia-Fatou theory of iteration of rational functions [10, 15] (and its subsequent developments) provides an insight into the dynamics of such iteration schemes. Nonconvergent sequences associated with Newton's method and its generalizations may, for example, exhibit asymptotic periodic or even chaotic behavior, neither of which may be simply explained away as results of calculations in finite precision arithmetic. As will be shown below, when the $g(z)$ constitute a one-parameter family of polynomials and the parameter is varied, the asymptotic behavior of nonconvergent sequences may

exhibit a cascade of period-doubling bifurcations eventually transforming to chaos, a characteristic feature of quadratic [7, 11, 16] and polynomial-like [8] maps.

The behavior of Newton's method as a deterministic dynamical system on the real line has been discussed in a recent article by Saari and Urenko [19]. Howland and Vaillancourt [13] investigated the pathological attractive cycles which may be encountered in Newton's method. Curry et al. [6], by means of a series of experiments and the main concepts of Julia-Fatou theory, catalogued the behavior patterns associated with Newton's method as applied to a one-parameter family of cubic polynomials. This present study was motivated in part by Ref. [6]. We consider a generalized family of Schröder iteration functions having the form

$$S_m(z) = z + \sum_{n=1}^{m-1} c_n [-g(z)]^n, \quad m = 2, 3, 4, \dots, \quad (1.2)$$

and constructed so that the iteration sequence $z_{n+1} = S_m(z_n)$ converges locally to a zero z^* of $g(z)$ as $O(|z - z^*|^m)$. The case $m = 2$ corresponds to the Newton method of Eq. (1.1). When $g(z)$ is a polynomial, the degree of the rational function $S_m(z)$ increases with m . For $m > 2$, the $S_m(z)$ functions may possess extra fixed points which are generally distinct from the roots z_i^* . If attractive, these points may trap the Schröder iteration sequence $\{z_n\}$.

The Schröder functions are discussed in Section 2. In Section 3 Newton's method and its generalizations are examined in the light of Julia-Fatou theory of iteration of rational functions. In Section 4 are presented microcomputer generated plots of the basins of attraction of the Schröder schemes for $m = 2$ and 3 as applied to the function $g(z) = z^4 - 1$. In each case the common boundary of these basins of attraction constitutes the Julia set of the Schröder iteration function. We also examine the dynamics of Schröder maps for a one-parameter family of cubic polynomials. There exist regions in complex parameter space where critical points of the $S_m(z)$ are attracted to points or cycles which do not correspond to roots of the $g_A(z)$. These regions exhibit the morphology and classical characteristics of Mandelbrot sets. In Section 5, another family of iteration functions of prescribed order, the König functions, is introduced and briefly examined.

2. The Schröder Iteration Functions

Let $f(z): \mathbb{C} \rightarrow \mathbb{C}$ be analytic on a compact subset T of the complex plane \mathbb{C} , having fixed point $p \in T$, i.e., $f(p) = p$. The fixed point p is *attractive*, *indifferent* or *repulsive* depending on whether $|f'(p)|$ is less than, equal to or greater than one. If $f'(p) = 0$, then p is termed *superattractive*. Given a starting value $z_0 \in T$, we define the iteration sequence $\{z_n\}_0^\infty$ by $z_{n+1} = f(z_n)$, $n = 0, 1, 2, \dots$. Now assume that p is attractive, i.e., that $z_n \rightarrow p$ as $n \rightarrow \infty$. Let $e_n = z_n - p$ be the error associated with the n th iterate. Using the Taylor expansion of $f(z)$, we have

$$\begin{aligned}
e_{n+1} &= z_{n+1} - p \\
&= f(e_n + p) - f(p) \\
&= \frac{1}{m!} f^{(m)}(p) (e_n)^m + O[(e_n)^m], \quad n \rightarrow \infty,
\end{aligned} \tag{2.1}$$

where m is the smallest integer for which $f^{(m)}(p) \neq 0$ (usually $m = 1$). Then $f(z)$ is said to be an *iteration function of order m* .

We now consider the construction of iteration functions of prescribed order to determine the simple roots of $g(z) = 0$. The first case, $m = 2$, corresponds to quadratic convergence and can yield the familiar Newton method. The point z^* is a zero of $g(z)$ iff it is a fixed point of $f(z) = z - h(z)g(z)$, where $h(z)$ is an arbitrary non-zero function analytic in T . The problem is to construct $f(z)$ so that $f'(z^*) = 0$. Since $f'(z) = 1 - h'(z)g(z) - h(z)g'(z)$ and $g(z^*) = 0$, we may choose $h(z) = [g'(z)]^{-1}$ to give Eq. (1.1) for $g'(z) \neq 0$. This represents a generalized and non-geometric procedure of constructing Newton's method over the complex plane.

Higher order iteration functions may be constructed in the same spirit. The *Schröder iteration functions* [20] have been defined in Eq. (1.2), with

$$c_n(z) = \frac{1}{n!} \left[\frac{1}{g'(z)} \frac{d}{dz} \right]^{n-1} \frac{1}{g'(z)}. \tag{2.2}$$

The coefficients $c_n(z)$ are analytic functions for $g'(z) \neq 0$. The iteration sequences defined by $z_{n+1} = S_m(z_n)$ converge locally to the zeros z_i^* of $g(z)$ as $O(|z - z_i^*|^m)$. To see this, we assume $g(z)$ to be analytic in T and $g'(z) \neq 0$. The functions $S_m(z)$ are analytic in T . For every $z^* \in T$ such that $g(z^*) = 0$, it follows that $S_m(z^*) = z^*$ and

$$S'_m(z^*) = S''_m(z^*) = \dots = S^{(m-1)}_m(z^*) = 0. \tag{2.3}$$

A proof of Eq. (2.3) is given in Henrici [12], p. 520.

The $S_m(z)$ functions in Eq. (1.2) are truncations of a general infinite series in $g(z)$, the first three terms of which are given explicitly below:

$$\begin{aligned}
S(z) = z - \frac{1}{g'(z)} g(z) - \frac{g''(z)}{2[g'(z)]^3} [g(z)]^2 \\
- \frac{\frac{1}{2} [g''(z)]^2 - \frac{1}{6} g'(z)g'''(z)}{[g'(z)]^5} [g(z)]^3 \dots
\end{aligned} \tag{2.4}$$

The construction of $S_m(z)$ requires a knowledge of the first $m-1$ derivatives of $g(z)$. From Eq. (1.2) we see that only for $m = 2$, the Newton method, does

the fixed point condition $S_m(p) = p$ imply that $g(p) = 0$. For $m > 2$, however, it implies that either (i) $g(p) = 0$ or (ii) $T_m(p) = 0$, where

$$T_m(z) = \sum_{n=0}^{m-2} c_{n+1}(z)[-g(z)]^n. \quad (2.5)$$

The introduction of these extra fixed points may complicate the root-finding procedure: as repulsive or indifferent fixed points, they alter the basins of attraction for the roots; as attractive fixed points, they may trap an iteration sequence.

3. Julia-Fatou Theory and Schröder Rational Iteration Functions

The theory of iteration of rational functions, originating in the classical research of Julia [15] and Fatou [10], has witnessed a dramatic resurgence of interest in the last twenty years. A comprehensive account of this research is given in the excellent review of Blanchard [3]. Details of many important proofs are given in the paper by Brolin [4]. Some important concepts and their connection with Newton-type iteration schemes are outlined below.

Let $R(z)$ be a rational function, $R(z) = P(z)/Q(z)$ where $P(z)$ and $Q(z)$ are polynomials with complex coefficients and no common factors, and $d = \deg(R) \equiv \max\{\deg(P), \deg(Q)\} \geq 2$. The sequence of iterates $\{R^n(z)\}$ of $R(z)$ is defined by

$$R^0(z) = z, \quad R^1(z) = R(z), \quad R^{n+1}(z) = R(R^n(z)), \quad n = 0, 1, 2, \dots$$

The inverses of $R(z)$ shall be denoted by $R_i^{-1}(z)$, where the subscript index $i = 1, 2, \dots, d$ enumerates all branches of the inverse. We now consider $R: \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$ where $\bar{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ denotes the Riemann sphere with suitably defined spherical metric. Given a point $z_0 \in \bar{\mathbb{C}}$, the iteration sequence $\{z_n\}_{n=0}^{\infty}$, given by

$$z_{n+1} = R(z_n) = R^{n+1}(z_0),$$

defines the *forward orbit* of z_0 .

If $R^k(p) = p$ and $R^m(p) \neq p$ for $m < k$, then p is a *fixed point of order* k . The set of distinct points $\{p_i, i = 1, 2, 3, \dots, k\}$, where

$$p_1 = R(p), \quad p_2 = R(p_1), \dots, p_k = R(p_{k-1}),$$

is termed a *k-cycle*. If $k = 1$, p is simply called a fixed point of $R(z)$. The *k-cycle* is *attractive*, *indifferent*, or *repulsive*, depending whether the multiplier $|[R^k(p_i)]'|$ is less than, equal to or greater than one, respectively.

The Julia set $J(R)$ of the rational map $R: \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$ is formally defined as the set of $z \in \bar{\mathbb{C}}$ for which the family of maps $R^n(z)$ is not normal, in

the sense of Montel [1]. A more working description is that $J(R)$ is the closure of all repulsive k -cycles of $R(z)$, $k = 1, 2, 3, \dots$. Its complement, $F = \bar{\mathbb{C}} \setminus J(R)$, the Fatou set, is the set of all $z \in \bar{\mathbb{C}}$ for which the family $R^n(z)$ is equicontinuous, in the spherical metric on some neighbourhood of each point of F .

Some important properties of $J(R)$ are listed below:

- (a) $J \neq \emptyset$ and J is closed;
- (b) J is invariant with respect to R , i.e., $R(J) = J = R^{-1}(J)$;
- (c) $J(R) = J(R^m)$, $m = 2, 3, 4, \dots$;
- (d) If J has interior points then $J = \bar{\mathbb{C}}$;
- (e) $J(R)$ is compact and non-denumerable. In general, its Hausdorff-Besicovitch dimension is non-integral, whereupon $J(R)$ is a fractal, as defined by Mandelbrot [17].

Let p be an attractive fixed point of $R(z)$. The *attractive basin* (stable set) $W(p)$ of p is defined as the set

$$W(p) = \{z \in \bar{\mathbb{C}} \mid R^n(z) \rightarrow p \text{ as } n \rightarrow \infty\}.$$

The *immediate attractive basin* $A(p)$ of p is the maximal domain containing p on which the sequence of iterates $\{R^n\}$ is normal. We now have the following important property: boundary of $W(p)$ is $J(R)$. It follows that if $R(z)$ has several distinct attractive points, then their basins of attraction share the same boundary, the Julia set of $R(z)$.

A simple and illustrative example is afforded by the map $R(z) = z^2$. The unit circle $C = \{z: |z| = 1\}$ is invariant with respect to $R(z)$ and its iterates. All fixed points of $R(z)$ and its iterates, except $z = 0$ and $z = \infty$ lie on C and are repulsive. C is the Julia set of $R(z)$. The forward orbit of any point given by $|z| < 1$ is the fixed point $z = 0$. The forward orbit of any point given by $|z| > 1$ is the point at infinity. The Julia set C may be regarded as a *repeller set* under the action of the forward map $R(z)$. Equivalently, C may be regarded as the *attractor set* for the inverses $R_1^{-1}(z) = +\sqrt{z}$, $R_2^{-1}(z) = -\sqrt{z}$.

Before the works of Julia and Fatou, Cayley [5] began an investigation of Newton's method in the complex plane. Firstly, he analyzed Newton's method of determining the square roots of unity, i.e., the zeroes of

$g(z) = z^2 - 1$, to ascertain the basin of attraction of each root. Here $N(z) = z/2 + 1/(2z)$. One would expect that the imaginary axis J is the boundary of the two attractive basins, $W(+1)$ and $W(-1)$, since it is easily shown that $N(J) = J$, i.e., if $z \in J$ then $N(z) \in J$. Now consider the conformal map $T(z) = (z-1)/(z+1)$. Let $R_+ = \{z: \operatorname{Re}(z) > 0\}$ and

$R_- = \{z: \operatorname{Re}(z) < 0\}$. Then $TJT^{-1} = C$ where C denotes the unit circle,

$TR_+T^{-1} = \{z: |z| < 1\}$, $TR_-T^{-1} = \{z: |z| > 1\}$ and $TNT^{-1} = z^2$. Furthermore,

$T(+1) = 0$ and $T(-1) = \infty$. $N(z)$ is conformally conjugate to the map $R(z) = z^2$

which was our earlier example. It now follows that under the action of $N(z)$, $W(+1) = R_+$, $W(-1) = R_-$ and J is the Julia set of $N(z)$. As such, J is the repeller set for $N(z)$ or the attractor set under the action of $N_i^{-1}(z)$, $i = 1, 2$. In the language of Barnsley and Demko [2], J is the attractor for the iterated function system (IFS) $\{T, w_+(z), w_-(z)\}$, where

$$\begin{aligned} w_+(z) &= z + \sqrt{z^2 - 1}, \\ w_-(z) &= z - \sqrt{z^2 - 1}. \end{aligned} \tag{3.1}$$

The situation is not so simple, as Cayley discovered [9, 18], for the analysis of Newton's method for higher roots of unity, i.e., $g_n(z) = z^n - 1$, $n > 2$. The Julia set J must serve as a common boundary for all $W(z_i^*)$. In other words, given any point $z \in J$, then any ε -neighborhood of z must include points from all the $W(z_i^*)$. To illustrate, Figure 1(a) presents the basins of attraction for Newton's method for $g_4(z) = z^4 - 1$. The Julia set boundary is a complicated curve which is infinitely self-similar -- a magnification of any region reveals further similar and intricate structure, characteristic of a fractal curve [17]. The Julia set includes points on the lines $\text{Re}(z) = \pm \text{Im}(z)$.

Figure 1(b) is an attractive basin map for the Schröder $S_3(z)$ method as applied to $g_4(z) = 0$, namely, $S_3(z) = (21z^8 + 14z^4 - 3)/(32z^7)$. The Julia set boundary is similar in structure to that of Figure 1(a) on the diagonals, but with a greater number of "petals" being nested in a self-similar fashion. A fundamental difference is noted in the immediate stable sets $A(z_i^*)$ of each root as they no longer extend into the origin as in Figure 1(a). This is due to the presence of four extra fixed points of $S_3(z)$, as given by the zeroes of Eq. (2.5). These fixed points are solutions of $z^4 - 3/11 = 0$. They are repulsive [21] and must lie on the Julia set $J(S_3)$. The appearance of these extra fixed points and their effects on the basins of attraction demonstrates the caution that may be necessary with the use of higher order iteration methods. The existence of repulsive fixed points for S_3 and S_4 iteration methods as applied to $g_n(z) = 0$ has been discussed for general n in [21].

4. Parameter Space and Chaotic Dynamics

As mentioned earlier, the Newton or Schröder methods are not guaranteed to converge to zero of $g(z)$. If the initial point $z_0 \in J$, the Julia set, then the sequence $\{z_n\}$ will always remain on J . Apart from this situation, however, the possibility exists that the z_n converge to periodic cycles or

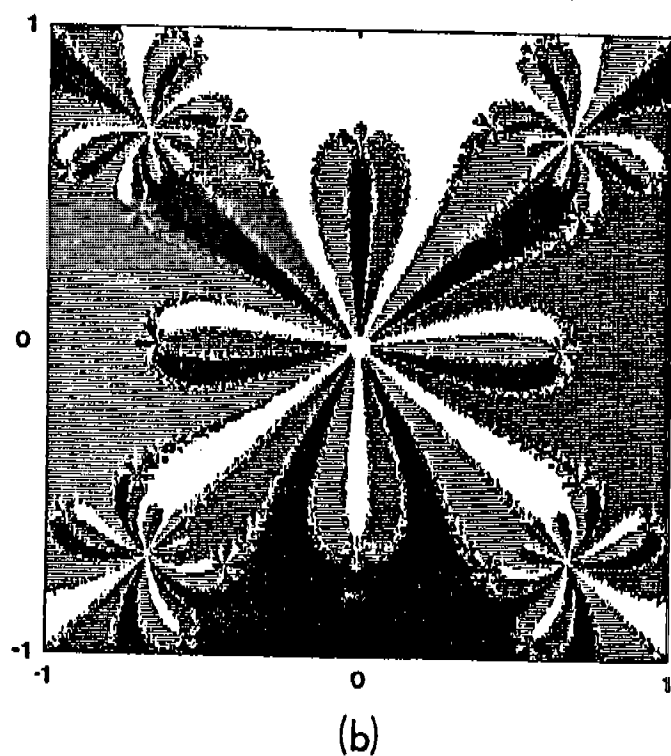
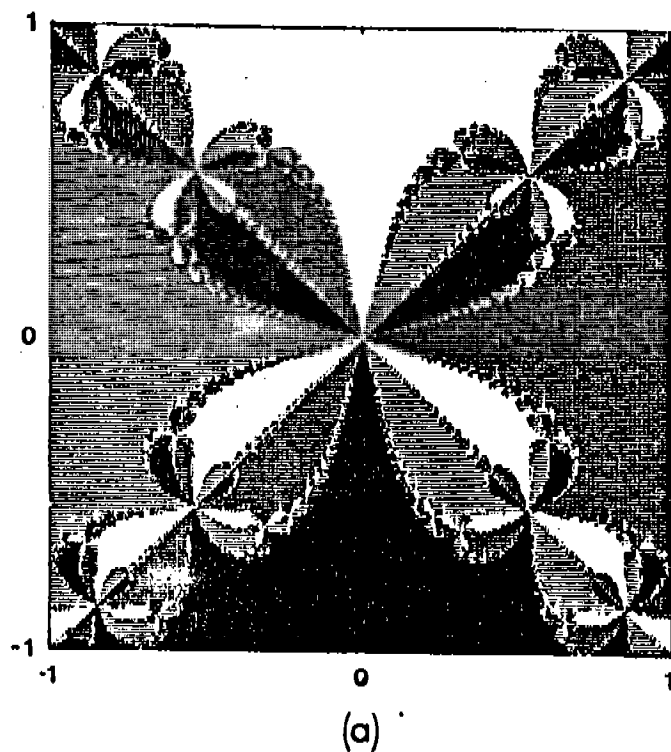


Figure 1. Schroder basins of attraction $W(z_i^*)$ for the roots of $z^4 - 1 = 0$ in the complex region $[-1, 1] \times [-1, 1]$. White regions constitute $W(i)$; black regions, $W(-i)$; light grey, $W(-1)$; dark grey, $W(1)$:
 (a) S_2 (Newton) method, (b) S_3 method.

exhibit chaotic behavior. Rather than trying to construct specific examples of such pathological behavior, we may systematically examine the iteration schemes associated with a parametrized family of polynomials $g(z)$. Here, we consider the one-parameter family of cubic polynomials

$$g_A(z) = z^3 + (A-1)z - A. \quad (4.1)$$

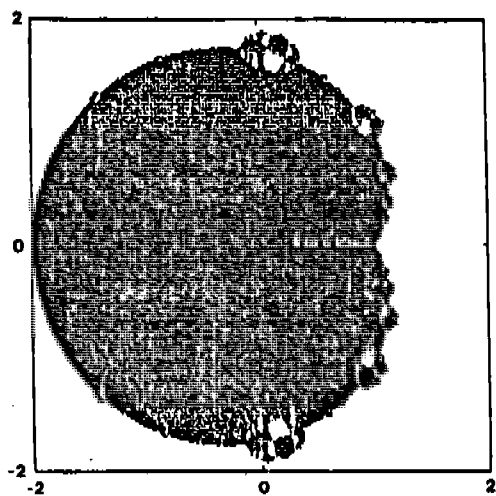
The zeros of $g_A(z)$ are $z_1^* = 1$, $z_{2,3}^* = (-1 \pm \sqrt{1-4A})/2$. We shall now be working in a parameter space where $A \in \mathbb{C}$. Each point $A = (\text{Re}(A), \text{Im}(A))$ represents a dynamical system with its own fixed points, Julia sets and possible attractive cycles.

Curry et al. [6] first examined Newton's method in this parameter space to discover regions in A -space where attractive periodic cycles exist. This feature is also observed for the higher order $S_m(z)$ functions as well as the possible existence of extra attractive fixed points corresponding to the roots of Eq. (2.4) [21]. Here we restrict our attention to the S_2 (Newton) and S_3 iteration schemes, the former of which will serve as a reference.

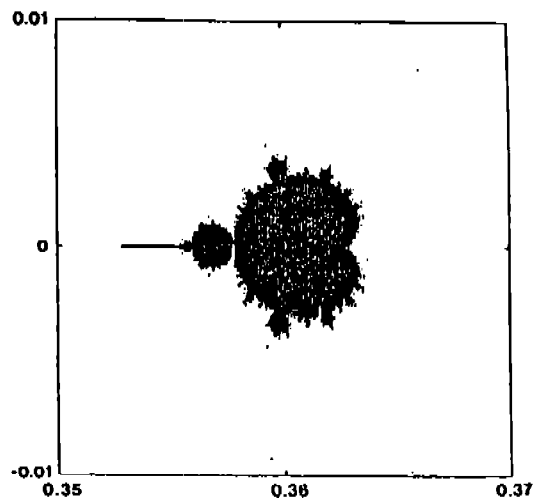
In order to detect the existence of attractive cycles which could interfere with the Schröder search for the z_1^* , we observe the orbits of the *critical points* of the $S_m(z)$, i.e., those points $c \in \mathbb{C}$ for which $S'_m(c) = 0$. The underlying reason for studying these special orbits rests in the following theorem of Fatou [10]: If $R(z)$ is a rational function having an attractive periodic cycle, then at least one critical point will converge to it.

Among the critical points of the $S_m(z)$ are the zeroes z_1^* which, of course, are also attractive fixed points (1-cycles) of the $S_m(z)$. These points are obviously not free to converge to any other attractive cycle. Other roots, which we shall call the *free critical points* c_i are available, however. The free critical points for the first two Schröder functions associated with the $g_A(z)$ are (i) for $S_2(z)$, $c_1 = 0$ and (ii) for $S_3(z)$, $c_{1,2} = \pm[(A-1)/15]^{1/2}$.

In order to study the dynamics of these maps on a microcomputer, a region of the complex A -plane was represented by a grid of 400×200 points, each point corresponding to a pixel of a computer video terminal. For each point $A = (\text{Re}(A), \text{Im}(A))$, a free critical point c_i was computed and used as an initial value for the iteration sequence, $z_{n+1} = S_m(z_n)$. After each iteration, the distance between the iterate z_k and each z_i^* was computed. If any of these distances was less than a prescribed value of (10^{-4}) the sequence was assumed to converge to that particular root and the corresponding



a



b

Figure 2. Parameter space maps associated with Newton's method for the one-parameter family of cubic polynomials $g_A(z)$.

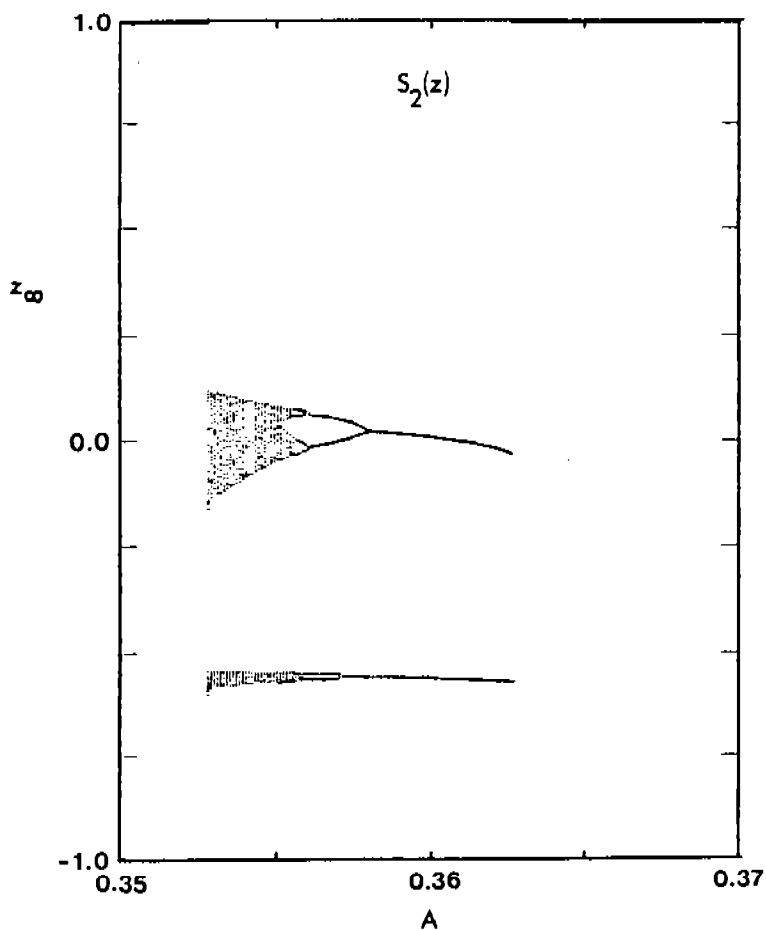


Figure 3. Asymptotic trajectories of the critical point c_1 for Newton's method as applied to the $g_A(z)$ in the range $0.35 \leq A \leq 0.37$.

pixel was colored accordingly. If no such convergence was observed after a prescribed number of iterations (typically 200), then the grid point was left black. The resulting black areas represented regions in parameter space for which additional cycles existed.

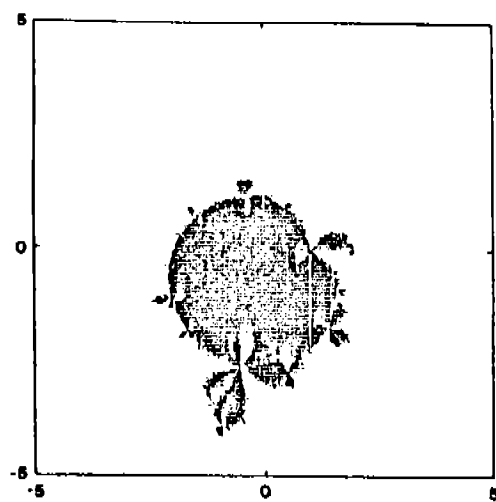
S₂-Newton Method

Figure 2(a) represents the region of the complex A-plane $[-2,2] \times [-2,2]$. White regions represent values of A for which the sole critical point $c_1 = 0$ is attracted to the root $z_1^* = 1$, grey regions correspond to attraction to either of the roots $z_{2,3}^*$. Small black areas, representing parameter values for which pathological attractive cycles exist, are observed at $A \cong (0.31, \pm 1.64)$ and $(1.01, \pm 0.98)$. When magnified, these regions have the same general shape as the remarkable Mandelbrot bifurcation sets [17] for quadratic maps $R(z) = z^2 - \lambda$. Four other sets are detected on the real axis at $A \cong 0.26, 0.36, 0.5$ and 0.65 . Figure 2(b), a magnification of the region $[0.35, 0.37] \times [0.01, 0.01]$, reveals a characteristic Mandelbrot-like set.

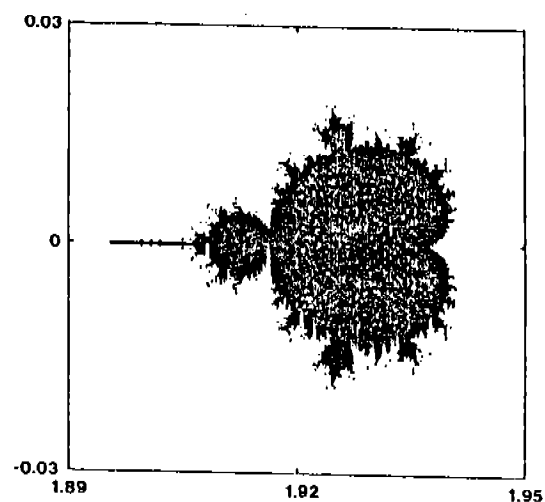
The existence of stable attractive periodic cycles and regions of period-doubling bifurcations corresponding to the region in Figure 2(b) is shown in Figure 3. Here, we plot the asymptotic trajectory (z_n for $n \geq 10000$) of the critical point c_1 for the range of real parameter values $0.35 \leq A \leq 0.37$. The calculations were performed on a CYBER 180/855 main-frame computer in double precision (32 significant digits). Let us examine the dynamics as the parameter A is decreased from 0.37. The free critical point c_1 is eventually mapped to $z_1^* = 1$. Below the critical value $A = 0.362683\dots$, c_1 is suddenly mapped into a 2-cycle. As A is further decreased the 2-cycle becomes a 4-cycle, etc. . The bifurcation to 8- and 16-cycles proceeds quite rapidly, with an eventual transition to chaotic behavior. A three-cycle is then observed, followed by a return to chaotic behavior, etc. . At $A \cong 0.35286$ a sudden return from chaos back to the fixed point $z_1^* = 1$ is observed. If the Mandelbrot set of Figure 2(b) is superimposed on Figure 3, it is seen that its pinch points correspond to the points at which the 2^n -cycles bifurcate.

S₃-Iteration Method

Figure 4(a) presents regions in parameter space $A \in [-5,5] \times [-5,5]$ for which the critical point c_1 is attracted either to $z_1^* = 1$ (white), $z_{2,3}^*$ (grey) or neither (black). The parameter space map for the other critical point $c_2 = -c_1$ is obtained from a reflection of the regions in this figure about the real A axis. An enlargement of the region $[1.89, 1.95] \times [-0.03, 0.03]$ is shown in Figure 4(b) along with its reflection about the real A-axis. The upper half (including real axis) of this Mandelbrot-like set corresponds to A values for which c_1 does not converge



a



b

Figure 4. Parameter space maps associated with the Schroder S_3 method as applied to the cubic polynomials $g_A(z)$.

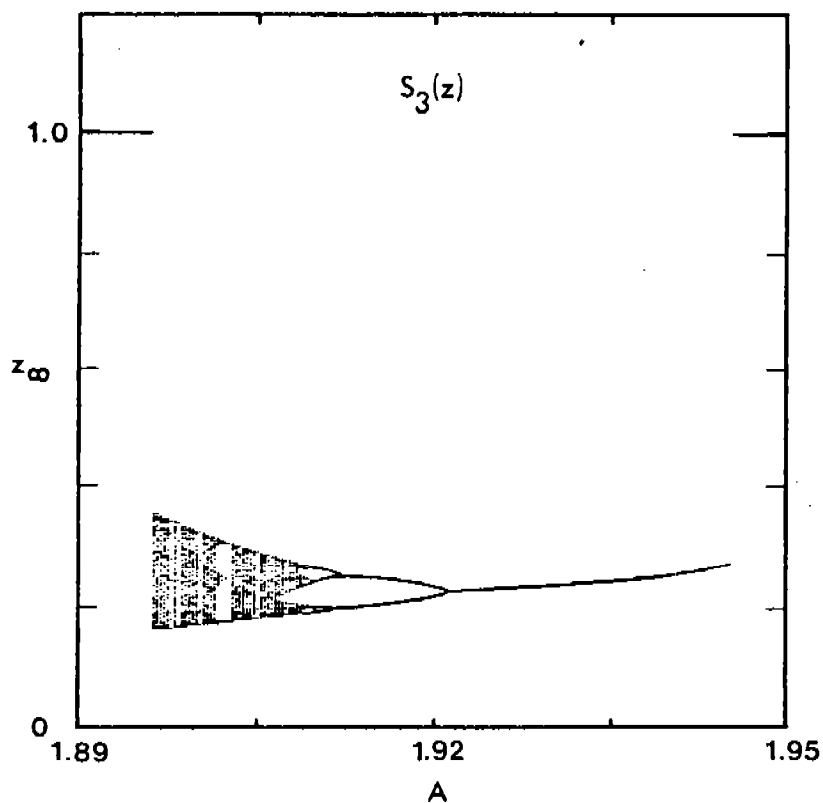


Figure 5. Asymptotic trajectories of the critical point c_1 for the S_3 method as applied to the $g_A(z)$ in the range $1.89 \leq A \leq 1.95$.

to the z_i^* ; the lower half corresponds to the critical point c_2 . Unlike in the Newton method, there are two possibilities for the orbits of the c_1 to be trapped away from the roots z_i^* :

- (i) convergence to a k -cycle as seen above,
- (ii) convergence to additional attractive fixed points given by the zeroes of Eq. (2.5), where $g(z) = g_A(z)$.

Figure 5 shows the asymptotic trajectories of the free critical point c_1 in the parameter region corresponding to Figure 4(b).

5. König Iteration Scheme

Here we briefly describe another set of iteration functions of prescribed order. Given a polynomial $g(z)$ with zeroes z_i^* , the König iteration functions are defined as [14]

$$K_m(z) = z + (m-1) \frac{[1/g(z)]^{(m-2)}}{[1/g(z)]^{(m-1)}}, \quad (5.1)$$

along with the iteration sequence defined by $z_{n+1} = K_m(z_n)$. If the z_n are sufficiently close to zero z_i^* of $g(z)$, then $z_n \rightarrow z_i^*$ as $O(|z_n - z_i^*|^m)$ [14]. The case $m = 2$ again corresponds to Newton's method.

Figure 6 shows the basins of attraction for the four roots of unity, i.e., $g(z) = z^4 - 1$, for the K_3 iteration procedure. There are major differences between this basin map and those associated with the S_2 (or K_2) and S_3 iteration methods of Figures 1(a) and 1(b). The immediate stable set of each root is much larger in Figure 6 since the "bubbly" Julia set regions are greatly compressed into the diagonals. The transition from the K_2 -Newton method to the K_3 method is significant. Secondly, the K_3 iteration function, as its S_3 counterpart, has four additional repulsive fixed points. Unlike the S_3 function, however, these points lie on the diagonal lines $\text{Re}(z) = \pm \text{Im}(z)$ and not on the real or imaginary axes. As such, they do not interfere with the immediate stable sets of the roots z_i^* .

We now consider the K_3 iteration method as applied to the one-parameter family of cubic polynomials $g_A(z)$ given in Eq. (4.1). Figure 7(a) represents the region of complex A -space $[-5, 5] \times [-5, 5]$. As before, regions are colored according to whether the free critical point $c_1 = [(A-1)/15]^{1/2}$ is mapped to $z_1^* = 1$ (white), $z_{2,3}^*$ (grey) or neither (black). Only one black region of nonconvergence, centered at $A \cong (1.99, -3.26)$ is detectable

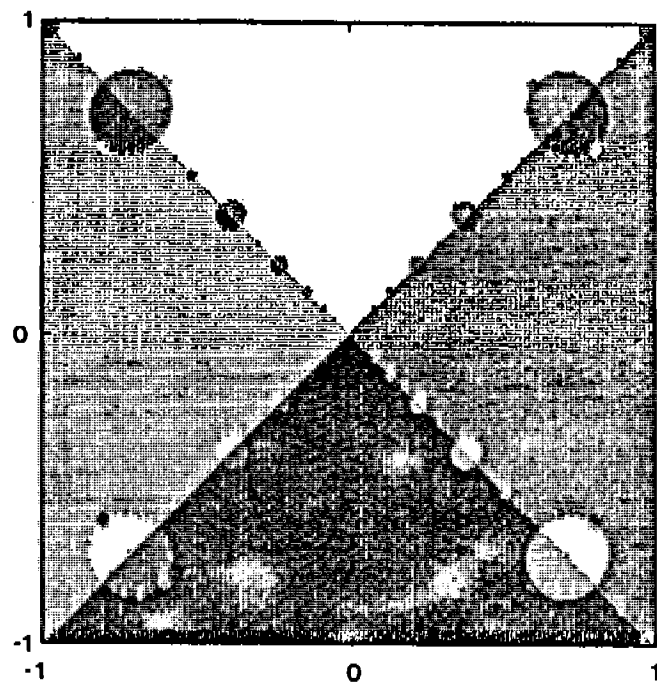
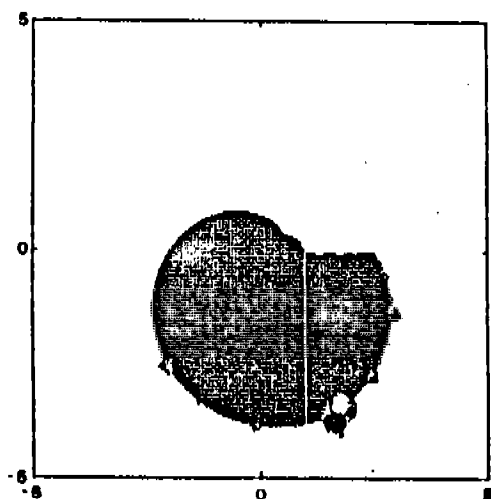
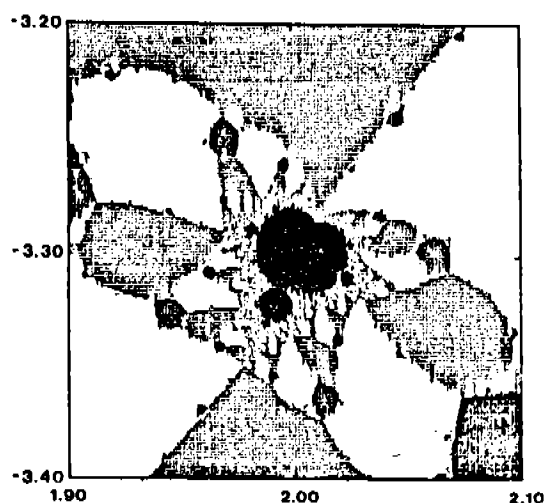


Figure 6. Basins of attraction $W(z_i^*)$ associated with the Konig K_3 iteration method as applied to $z^4 - 1 = 0$. Regions are shaded as in Fig. 1.



a



b

Figure 7. Parameter space maps associated with the Konig K_3 method as applied to the cubic polynomial family $g_A(z)$.

in this plot. A magnification of this region is presented in Figure 7(b) to reveal a Mandelbrot-like set. Trajectories of c_1 for sample parameter values lying in this set have been calculated. Within the largest cardioid-like region of the Mandelbrot set, c_1 is mapped to 2-cycles. As the parameter A is varied along the axis of symmetry a period-doubling cascade eventually leading to chaotic behavior is observed as in Section 4 for the Schröder iteration methods. Interestingly, in no regions of the K_3 Mandelbrot set are the c_1 observed to map to fixed points. In fact it can be shown [22] that there exist no regions in the K_3 parameter space where attractive fixed points other than the z_i^* exist.

Acknowledgements

The author wishes to thank Prof. M. F. Barnsley for stimulating and informative discussions during the course of these investigations. The support of a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship for 1984 and 1985 is gratefully acknowledged. The pictures presented in this report were computed on a Sanyo MBC-555-2 micro-computer and printed on an Epson FX-80 dot-matrix printer.

References

1. L. Ahlfors, Complex Analysis, McGraw-Hill, New York, 1979.
2. M. F. Barnsley and S. Demko, "Iterated function systems and the global construction of fractals," to appear in Proc. Roy. Soc. London Ser. A.
3. P. Blanchard, "Complex analytic dynamics on the Riemann sphere," Bull. Amer. Math. Soc., v. 11, 1984, pp. 85-141.
4. H. Brolin, "Invariant sets under iteration of rational functions," Ark. Math., v. 6, 1966, pp. 103-144.
5. A. Cayley, "Application of the Newton-Fourier method to an imaginary root of an equation," Quart. J. Pure Appl. Math., v. 16, 1879, pp. 179-185; "Sur les racines d'une équation algébrique," CRAS, v. 110, 1890, pp. 215-218.
6. J. H. Curry, L. Garnett and D. Sullivan, "On the iteration of a rational function: computer experiments with Newton's method," Commun. Math. Phys., v. 91, 1983, pp. 267-277.
7. A. Douady and J. Hubbard, "Itération des polynômes quadratiques complexes," C.R. Acad. Sc. Paris, v. 294, 1982, pp. 123-126.
8. A. Douady and J. Hubbard, "On the dynamics of polynomial-like mappings," 1984, preprint.

9. J. P. Eckmann, "Savez-vous résoudre z^3-1 ?" La Recherche, v. 14, 1983, pp. 260-262.
10. P. Fatou, "Sur les équations fonctionnelles," Bull. Soc. Math. France, v. 47, 1919, pp. 161-271; v. 48, 1920, pp. 33-94, 208-314.
11. M. Feigenbaum, "Quantitative universality for a class of nonlinear transformations," J. Stat. Phys., v. 19, 1978, pp. 25-52.
12. P. Henrici, Applied and Computational Complex Analysis, v. 1, Wiley, New York, 1974.
13. J. L. Howland and R. Vaillancourt, "Attractive cycles in the iteration of meromorphic functions," 1984, preprint.
14. A. S. Householder, Principles of Numerical Analysis, Mc-Graw Hill, New York, 1953.
15. G. Julia, "Mémoire sur l'itération des fonctions rationnelles," J. Math. Pures Appl., v. 4, 1918, pp. 47-245.
16. B. Mandelbrot, "Fractal aspects of $z \rightarrow \lambda z(1-z)$ for complex λ and z ," Ann. N.Y. Acad. Sci., v. 357, 1985, pp. 249-259.
17. B. Mandelbrot, The Fractal Geometry of Nature, W. H. Freeman, New York, 1983.
18. H. O. Peitgen, D. Saupe and F. V. Haeseler, "Cayley's problem and Julia sets," Math. Intelligencer, v. 6, 1984, pp. 11-20.
19. D. G. Saari and J. B. Urenko, "Newton's method, circle maps, and chaotic motion," Amer. Math. Monthly, v. 91, 1984, pp. 3-17.
20. E. Schröder, "Ueber unendlich viele Algorithmen zur Auflösung der Gleichungen," Math. Ann., v. 2, 1870, pp. 317-364.
21. E. R. Vrscay, "Julia sets and Mandelbrot-like sets associated with higher order Schroder rational iteration functions," to appear in Math. Comp.
22. E. R. Vrscay, "Julia sets and chaotic dynamics associated with König rational iteration functions," (unpublished).

CHAOTIC EIGENSTATES FOR QUANTUM MECHANICAL SYSTEMS

D. Bessis

SACLAY, France and School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT. We shall first discuss algebraic properties of the iterations of polynomials and show that these properties are related to almost periodic Schrödinger operators. An exactly solvable model is introduced which displays interesting features: almost periodicity, singular spectrum, chaotic states, exact renormalization group.

I. INTRODUCTION. Singular continuous spectra arise in many different physical models, connected with either fractal structures or almost periodic potentials in Schrödinger equations. Physical problems in which scaling properties play an important role lead to a detailed interpretation of the observed phenomena in terms of fractal structures [1], such as vibration properties of proteins [2], percolation in discontinuous thin films [3] and diamagnetic properties of superconductors near the percolation threshold [4]. Also, models involving a fractal structure as underlying framework generate singular measures in a natural way, as for instance the vibration spectrum on Sierpinsky's gasket [5] or Potts models on hierarchical lattices [6].

Another important class of physical models which generate singular continuous spectra is that of the almost periodic Schrödinger operators [7]. They appear for instance in incommensurate structures, conducting or superconducting linear chains [8], in the electronic properties of crystals in a magnetic field [9] or more generally in metal-insulator transitions, the almost Mathieu equation being an example [10]. Very little is known about the nature and behavior of the wave functions for states belonging to a singular continuous spectrum.

Many discretized equations can be considered as Poincaré maps of dynamical systems with an infinite number of degrees of freedom, an analogy which has been used in the Frenkel-Kontorova model [11]. In this article we shall summarize the relation between one of the simplest dynamical systems, the iteration of polynomial mappings, and Schrödinger operators.

In Section II, we introduce an invariant measure under a polynomial transformation. In Section III we show orthogonality properties of iterated polynomials, and associate to those a Hilbert space operator the spectrum of which is

the Julia set of the polynomial transformation. In Section IV we introduce a one dimensional discrete Schrödinger operator the density of states of which is the invariant measure. In Section V we show that this operator has almost periodic properties and that the corresponding eigenstates display chaotic behavior.

II. INVARIANT MEASURES UNDER POLYNOMIAL TRANSFORMATIONS.

Let us consider a polynomial $T(x)$ of degree d written in canonical form

$$T(x) \equiv x^d + a_1 x^{d-1} + a_2 x^{d-2} + \dots + a_{d-1} x + a_d. \quad (\text{II.1})$$

We further assume that there exists a finite interval S of the real line such that for any $x \in S$, all the roots of the equation $T(y) = x$ are real and belong to S . This condition imposes weak constraints on the real coefficients a_2, a_3, \dots, a_d .

Consider an arbitrary point x_0 in S . Let $x_{(1)}^i = T_i^{(-1)}(x_0)$ $i = 1, 2, \dots, d$ be the d first preimages of x_0 , that is, the d different points which are mapped to x_0 by T . More generally, let $T^{(n)}$ be the n th iterate of T

$$T^{(1)}(x) = T(x), \quad (\text{II.2})$$

$$T^{(n)}(x) = T^{(n-1)}[T(x)],$$

and let $x_{(n)}^i = T_i^{(-n)}(x_0)$, $i = 1, 2, \dots, d^n$ be the d^n roots of the equation $T^{(n)}(x) = x_0$. The set of accumulation points of all preimages $x_{(n)}^i$ of x_0 is the Julia set of the polynomial T [12,13,14]. Under our hypothesis this set is real, contained in S , and independent of x_0 .

Following Brolin [14], we shall consider the asymptotic distribution of the predecessors: we define, for any n and for an arbitrary (but fixed) x_0 , the measure

$$d\mu_n(x) = \frac{1}{d^n} \left\{ \sum_{i=1}^{d^n} \delta(x - x_{(n)}^i) \right\} dx. \quad (\text{II.3})$$

This is a discrete measure with equal weights on all the preimages of order n of x_0 . Brolin [14] asserts that the sequence $d\mu_n$ has a limit in the weak topology, which is independent of x_0 , when n goes to infinity. This limiting measure $d\mu(x)$ has been recognized to have special invariance

properties; it is invariant under T , and gives equal weight to all inverse branches of T . This is sometimes referred to as the balanced property [15].

For our purpose here, it is sufficient to define the complexification of the above defined measure, which is supported by the Julia set contained in S . We define

$$G(z) = \int_S \frac{d\mu(x)}{z-s}. \quad (\text{II.4})$$

The invariance property is reflected by the following functional equation:

$$G(z) = \frac{1}{d} T'(z) G(T(z)). \quad (\text{II.5})$$

Expanding $G(z)$ around $z = \infty$, we get:

$$G(z) = \sum_{n=0}^{+\infty} \frac{\mu_n}{z^{n+1}}, \quad (\text{II.6})$$

where the μ_n are the moments of the measure $d\mu(x)$

$$\mu_n = \int_S x^n d\mu(x). \quad (\text{II.7})$$

It is easy to see that (II.5) allows one to compute the moments μ_n recursively, provided one normalizes μ_0 to the value +1.

The invariance properties of the measure are best summarized by the following identity:

$$\int_S \phi[T(x), x] d\mu(x) = \frac{1}{d} \int_S \left\{ \sum_{i=1}^d \phi(x, T_i^{-1}(x)) \right\} d\mu(x), \quad (\text{II.8})$$

where ϕ is an arbitrary measurable function of two variables.

III. ORTHOGONAL POLYNOMIALS AND THE HILBERT SPACE OPERATOR ASSOCIATED TO A JULIA SET. It is natural to introduce the set of orthogonal polynomials associated to the positive measure $d\mu(x)$. We consider the set of polynomials $P_n(x)$ of degree $n = 0, 1, 2, \dots, \infty$ with highest degree coefficient equal to 1, which satisfy

$$\int_S P_m(x) P_n(x) d\mu(x) = h_n \delta_{m,n}. \quad (\text{III.1})$$

Using (II.8) in an appropriate way, one finds [15,16,17,18, 19] that

$$P_n(T(x)) = P_{nd}(x). \quad (\text{III.2})$$

Iterating (III.2), one gets

$$P_n(T^{(k)}(x)) = P_{nd^k}(x). \quad (\text{III.3})$$

For $n = 1$ (III.3) reduces to

$$P_1(T^{(k)}(x)) = P_{d^k}(x). \quad (\text{III.4})$$

However, it can be checked that $P_1(x) \equiv x$ when $T(x)$ is written in canonical form. Therefore

$$T^{(k)}(x) \equiv P_{d^k}(x). \quad (\text{III.5})$$

This is a very remarkable result, because it explicitly states that the iterates of any polynomial (in canonical form) are subsets of the family of orthogonal polynomials with respect to the equilibrium measure associated to the Julia set corresponding to this polynomial. While the iteration of polynomials is a very complicated nonlinear operation, orthogonal polynomials satisfy a three-term linear recursive relation [20]:

$$P_{n+1}(x) = (x - \alpha_n)P_n(x) - R_n P_{n-1}(x). \quad (\text{III.6})$$

It is not difficult to obtain the α_n and R_n explicitly in terms of the coefficients a_2, a_3, \dots, a_d [19]. Therefore the nonlinear substitution of a polynomial into a polynomial has been changed into a linear operation. However, the linear relation (III.6) involves all the interpolating polynomials between the polynomials of the subfamily of degree d^k which are the k th iterates of $T(x)$. Nevertheless extremely interesting new points of view can be derived from (III.6).

Introducing the Jacobi matrix H associated to the three-term recursive relation (III.6) as well as a decimation operator D , acting on the infinite-dimensional vector with components

$$\psi_n(x) = \bar{P}_n(x) = h_n^{-1/2} P_n(x), \quad (\text{III.7})$$

where $\bar{P}_n(x)$ are the set of orthonormalized polynomials, one gets [21, 22]

$$HD = D(T(H)) \quad (\text{III.8})$$

where D is defined by

$$D\psi_n(x) = \psi_{dn}(x) \quad (\text{III.9})$$

and

$$H\psi(x) = x\psi(x). \quad (\text{III.10})$$

(III.10) expresses nothing other than the content of the three-term recursive relation (III.6).

However, if ψ is an eigenstate of H with eigenvalue x , (III.8) tells us that $D\psi$ is an eigenstate of H with eigenvalue $T(x)$, and therefore the spectrum of H is invariant under $T(x)$. That is, it is the Julia set associated with $T(x)$. It can be shown also that the spectrum is invariant under the inverse map $T^{-1}(x)$.

To conclude this first part, we see that, to any polynomial map, one can associate a Hilbert space operator the spectrum of which is the Julia set corresponding to this map, and whose eigenstate is an infinite dimensional vector the components of which are nothing but the set of orthogonal polynomials with respect to the equilibrium measure defined on the Julia set.

IV. ONE-DIMENSIONAL SCHRÖDINGER OPERATOR ASSOCIATED TO A POLYNOMIAL TRANSFORMATION. For the sake of simplicity we shall confine ourselves to the logistic map

$$T(x) = x^2 - \lambda. \quad (\text{IV.1})$$

We require $\lambda > 2$ for the Julia set to be real. In that case the Jacobi matrix H is an infinite tridiagonal matrix with all elements zero, except

$$H_{j,j+1} = H_{j+1,j} = \sqrt{R_{j+1}} \quad j = 0, 1, 2, \dots \quad (\text{IV.2})$$

and the $R_n(\lambda)$ are rational functions of λ given by the recurrence [17]

$$\begin{aligned} R_0 &= 0 \\ R_{2n} + R_{2n+1} &= \lambda \\ R_{2n}R_{2n-1} &= R_n \end{aligned} \quad (\text{IV.3})$$

which fits for the first few: $R_1 = \lambda$; $R_2 = 1$; $R_3 = \lambda - 1$; $R_4 = \frac{1}{\lambda - 1}$; $R_5 = \frac{\lambda^2 - \lambda - 1}{\lambda - 1}$; ...

The Schrödinger operator associated to the Jacobi matrix

H, has the following properties [21]

(i) Its spectrum is invariant under both

$$T(x) = x^2 - \lambda \quad (\text{IV.4})$$

and its two inverses

$$T^{-1}(x) = \pm \sqrt{x + \lambda}. \quad (\text{IV.5})$$

(ii) The integrated density of states is the equilibrium measure $d\mu(x)$.

(iii) When $\lambda > 2$, the spectrum of H is the set K of points $x(\vec{\sigma})$ where

$$\vec{\sigma} = (\sigma_0, \sigma_1, \dots, \sigma_n, \dots) \quad \sigma_i = \pm 1 \quad (\text{IV.6})$$

$$x(\vec{\sigma}) = \sigma_0 + \sqrt{\lambda + \sigma_1 \sqrt{\lambda + \sigma_2 \sqrt{\lambda + \dots}}} \quad (\text{IV.7})$$

(iv) K is a Cantor set of Lebesgue measure zero [14].

(v) The representation (IV.6), (IV.7) is a well adapted coding of K and the action of T is expressed on the sequences of signs $\vec{\sigma}$ as the usual shift S

$$\begin{aligned} S(\sigma_0, \sigma_1, \dots) &= (\sigma_1, \sigma_2, \dots) \\ T(x(\vec{\sigma})) &= x(S\vec{\sigma}). \end{aligned} \quad (\text{IV.8})$$

Similarly for the action of T^{-1}

$$\begin{aligned} \vec{\sigma}_{\pm} &= (\pm 1, \sigma_0, \sigma_1, \dots) \\ T^{-1}(x(\vec{\sigma})) &= x(\vec{\sigma}_{\pm}). \end{aligned} \quad (\text{IV.9})$$

Using the coding, one can identify the measure $d\mu(x)$ as the coin-tossing probability measure

$$\int f(x) d\mu(x) = \int \prod_{n=0}^{\infty} \left\{ d\sigma_n \frac{1}{2} [\delta(\sigma_n - 1) + \delta(\sigma_n + 1)] \right\} f(x(\sigma_0, \sigma_1, \sigma_2, \dots)). \quad (\text{IV.10})$$

Therefore we see that $d\mu$ has no atomic part and the action of T on the spectrum has the ergodic properties of a Bernoulli shift.

V. ALMOST PERIODICITY AND BEHAVIOR OF THE EIGENSTATE.

We come now to one of the most remarkable and unexpected facts about the coefficients R_n appearing in the Jacobi matrix H . A careful analysis of the recursion relations (IV.3) allows one to prove [17] that for $\lambda > 2$

$$\begin{aligned} 0 < R_{2n} &\leq 1 \\ \lambda - 1 &\leq R_{2n+1} < \lambda \\ \lim_{k \rightarrow \infty} R_{p2^k+s} &= R_s. \end{aligned} \tag{V.1}$$

Also, that for $\lambda \geq 3$,

$$|R_{p2^k+s} - R_s| \leq \frac{\lambda}{(\lambda-2)^k}, \tag{V.2}$$

which shows that the sequence R_n is almost periodic [22].

Therefore one can expand R_n in Fourier-like series

$$R_n = \sum_{q=0}^{\infty} \sum_{p=0}^{2^q-1} r_{p,q} \exp \left\{ \frac{2in\pi (2p+1)}{2^q} \right\}. \tag{V.3}$$

Those properties can be extended to complex values of λ large enough and for λ real and slightly bigger than 2 [23]. Those quasi-periodic properties also extend to the most general polynomial $T(x)$ of degree d .

Let us end this section by mentioning some properties of the states $\psi_n(x)$.

(i) Outside the spectrum, $\psi_n(x)$ increases exponentially, because $\psi_{n2^k}(x) = \psi_n(T^{(k)}(x))$ and $T^{(k)}(x)$ goes to infinity as x^{2^k} when x is outside the spectrum.

(ii) Inside the spectrum one gets the bound

$$|\psi_n(x)| < \frac{1}{2} \left(\frac{4\sqrt{\lambda}}{\sqrt{1+4\lambda}-1} \right)^k \quad \text{for } 2^{k-1} \leq n < 2^k. \tag{V.4}$$

Therefore we have an explicit polynomial bound. The Lyapunov exponent [24] can be proven to satisfy

$$2\gamma(x) = \gamma(T(x)), \tag{V.5}$$

which proves that $\gamma(x)$ vanishes on the spectrum, consistent

with the bound (V.4).

Finally, let us emphasize the following fact. On the spectrum we have

$$\psi_{p2^k}(x) = \psi_p(T^{(k)}(x)) . \quad (V.6)$$

However, when k goes to infinity, the sequence $T^{(k)}(x)$ is ergodic on the spectrum. Therefore the sequences $\psi_{p2^k}(x)$ have fully chaotic behavior of a precise type related to the Bernoulli shift mentioned above. Although the complete behavior remains to be analyzed [25], and it should not be excluded a priori that the chaotic behavior could be attributed to the sampling $(p2^k)$, we assert that ours is the only almost periodic discrete model in which such an explicit statement on the states can be made when the spectrum is singular continuous.

VI. CONCLUSION. A most fascinating point is to study the time dependent behavior of the corresponding quantum mechanical system

$$\psi(t) = e^{iHt}\psi(0), \quad (VI.1)$$

choosing as $\psi(0)$ the vector $(1, 0, 0, \dots) = |0\rangle$. The projection of $\psi(t)$ on $\psi(0)$:

$$\langle \psi(0), e^{iHt}\psi(0) \rangle = \int e^{ixt} d\mu(x). \quad (VI.2)$$

Therefore the Fourier transform of the measure which for large time analyzes the structure of the fractal on which it is defined [26] is likely to give a sequence of increasing times $t_1, t_2, t_3, \dots, t_n, \dots$ for which, no matter how large t is chosen, the probability that the system is found in its initial state remains finite [27]. Such singular spectra will provide systems which are intermediate between bound systems and unbound systems.

ACKNOWLEDGMENTS. I wish to thank Prof. Evans Harrell for a careful reading of the manuscript and Prof. William Ames for his kind invitation to the Third Army Conference.

References

- [1] B. Mandelbrot, "The Fractal Geometry of Nature," W. H. Freeman, San Francisco, 1982.
- [2] H. J. Stapleton, J. P. Allen, C. P. Flynn, D. G. Stinton, S. R. Kurtz, "The fractal form of proteins," Phys. Rev. Lett. 45 (1980), 1456.
- [3] R. F. Voss, R. B. Laibowitz, and E. I. Alessandrini, "Superconducting diamagnetism near the percolation threshold," J. Phys. Lett. 44 (1983), L-65.
- [4] R. Rammal, T. C. Lubensky, G. Toulouse, J. de Phys. Lett. 44 (1983), L-65.
- [5] E. Domany, S. Alexander, D. Ben Simon and L. P. Kadanoff, "Solutions to the Schrödinger equation on some fractal lattices," Phys. Rev. B 28 (1983), 3110.
- [6] B. Derrida, L. De Seze, C. Itzykson, "Fractal structures of zeros in hierarchical lattices," J. Stat. Phys. 3 (1983), 559.
- [7] D. R. Hofstadter, "Energy levels and wave functions of Bloch electrons in a rational and irrational magnetic field," Phys. Rev. B 16 (1967), 2239.
- [8] T. A. Turkevitch and R. A. Klemin, "Ginzburg-Landau theory of the upper critical field in filamentary superconductors," Phys. Rev. B. 19 (1979), 2520.
- [9] P. G. Harper, "Single band motion of conduction electrons in a uniform magnetic field," Proc. Phys. Soc. A 68 (1980), 874.
M. Ya. Azbel, "Energy spectrum of conductivity electrons in a magnetic field," Sov. Phys. JETP 19 (1964), 634.
- [10] J. Avron and B. Simon, "Singular continuous spectrum for a class of almost periodic Jacobi matrices," Bull. AMS 6 (1982), 81.
- [11] S. Aubry and P. V. Ledaeron, "The discrete Frenkel-Kontorova model and its extensions," CEN-SACLAY Preprint.
- [12] G. Julia, J. de Math. Ser. 7 (Paris), 47-245 (1918).
P. Fatou, Bull. Soc. Math. France, 47, 161-271 (1919); 48, 33-94 (1920); 48, 208-314 (1920).
- [13] A. Douady, Systèmes Dynamiques Holomorphes, Seminaire Bourbaki n° 599, November 1982.

- [14] H. Brolin, Ark. Mat. 6, 103-144 (1965).
- [15] M. F. Barnsley, J. S. Geronimo, A. N. Harrington, Bull. Amer. Math. Soc. 7, 381-384 (1982).
- [16] D. Bessis, M. L. Mehta, P. Moussa, C. R. Acad. Sci. Paris, 293, Ser. 1, 705-708 (1981).
- [17] D. Bessis, M. L. Mehta, P. Moussa, Letters Math. Phys. 6, 123-140 (1982).
- [18] M. F. Barnsley, J. S. Geronimo, A. N. Harrington, Commun. Math. Phys. 88, 479-501 (1983).
- [19] D. Bessis, P. Moussa, Commun. Math. Phys. 88, 503-529 (1983).
- [20] G. Szegő, Orthogonal Polynomials, Amer. Math. Soc. Colloquium publication, 23 (1939).
- [21] J. Bellissard, D. Bessis, P. Moussa, Phys. Rev. Lett. 49, 701-704 (1982).
- [22] H. Bohr, Almost periodic functions, Chelsea, New York (1951).
- [23] G. A. Baker, D. Bessis, P. Moussa, VIIth Conference on Mathematics and Physics, Boulder, Colorado, August 1983.
- [24] D. J. Thouless, J. of Phys. C, 5, 77-81 (1972).
- [25] M. Kohmoto, Y. Oono, Cantor spectrum for an almost periodic Schrödinger equation and a dynamical map, Illinois University at Urbana Preprint (1983).
- [26] D. Bessis, et al. to be published.
- [27] A. Zygmund, Trigonometric Series, Cambridge University Press, 1968, Vol. I, 194.

LARGE DEFORMATIONS OF ELASTOMER CYLINDERS SUBJECTED TO END
THRUST AND PROBE PENETRATION

A. R. Johnson*, C. J. Quigley*, and I. Fried**

Army Materials and Mechanics Research Center (AMMRC)
(AMXMR - SMM)
Watertown, MA 02172 - 2719

ABSTRACT. The Army is currently evaluating new elastomer tank track pads. These pads are repeatedly loaded to large strains when the tank is traveling on paved highways and to even larger strains when the tank is off the road and the pad is penetrated by sharp objects in the soil. In the process of evaluating the new elastomers, laboratory tests are made of cylindrical samples. In this effort a numerical method for analyzing tests of the cylindrical samples is presented. Axisymmetric triangular finite elements are used in the analysis to discretize the potential energy of the elastomer. Gradient and tangent matrices for the discretized potential energy are computed. The formulation allows different nonlinear energy density functionals to be used. Both end thrust and probe penetration problems are solved. A penalty method is used for the (no friction) contact problem associated with the probe penetration. Contours of the principal stretch ratios and stresses are shown on the deformed meshes for both the end thrust and probe penetration problems.

INTRODUCTION. The analysis of large deformations of elastomers involves joining currently active research work in numerical analysis and nonlinear mechanics[1-5]. In this effort we demonstrate a new numerical method for analyzing the large deformations of axisymmetric elastomers. The geometrical relations necessary for describing the stretch ratios of the deformed body in terms of the undeformed geometry are given. Both the deformed and undeformed bodies are interpolated using the finite element method and relationships between these interpolations are used to obtain the approximations to the variables which describe the stretch ratios. The internal energy of the deformed elastomer is expressed as the sum of separate symmetrical functions of the principal stretch ratios, see Valanis and Landel [6]. The internal energy is then used to determine the potential energy of the deformed elastomer. The energy gradient and tangent matrices with respect to the finite element nodal variables are then determined so that the Newton - Raphson method may be employed to obtain the deformed geometry (i.e., a minimum of the

* Mechanical Engineer, AMMRC

** Professor of Mathematics, Boston University

potential energy functional). The incompressibility constraint is enforced by adding the work done by changing the element's volume to the potential energy. End thrust is analyzed by enforcing the displacement of the top boundary. The probe penetration problem is approximated by adding a penalty term to the potential energy in which the distance between the contact surface (probe surface) and the nodes which penetrate into the contact region is minimized [5].

STRETCH RATIOS IN AXISYMMETRIC SOLIDS. The strain energy in a deformed elastomer solid can be determined in terms of the changes in length of material lines originating in the undeformed body [8]. The measure of this change in length most commonly used is the stretch ratio, defined as the ratio of the deformed material line length to the undeformed length. The stretch ratios in an axisymmetric solid can be determined as follows, see Figure 1. Let the coordinate system of the undeformed body be the (α, β) system and the system of the deformed body be the (r, z) system where the radial direction is given by the α, r coordinates and the axial direction by the β, z coordinates. Neighboring points in the undeformed body are located by the coordinates (α, β) and $(\alpha + d\alpha, \beta + d\beta)$ respectively. Deforming the elastomer results in the above points being mapped to (r, z) and $(r + dr, z + dz)$, respectively, in the deformed body. The line segments $\bar{\delta}_1$ and $\bar{\delta}_2$ represent the undeformed and associated deformed material lines for the direction θ at (α, β) in the undeformed body. If we let s be a coordinate line measured along $\bar{\delta}_1$ and assume that the (r, z) coordinates along $\bar{\delta}_2$ can be described by functions of s then we can write $\bar{\delta}_1$ and $\bar{\delta}_2$ as follows.

$$\begin{aligned}\bar{\delta}_1(\theta) &= d\alpha\hat{\alpha} + d\beta\hat{\beta} = (\cos\theta\hat{\alpha} + \sin\theta\hat{\beta})ds \\ \bar{\delta}_2(\theta) &= dr\hat{\alpha} + dz\hat{\beta} = (r_s\hat{\alpha} + z_s\hat{\beta})ds\end{aligned}\tag{1}$$

We can now determine the stretch ratio at (α, β) in the direction θ . It is

$$\lambda(\theta) = |\bar{\delta}_2| / |\bar{\delta}_1| = (r_s^2 + z_s^2)^{1/2}\tag{2}$$

To develop the finite element, we determine the principal stretch ratios in terms of the deformed and undeformed coordinates. We have the stretch ratio at (α, β) in direction θ using the deformed coordinates and the auxiliary system (θ, s) . We map the (θ, s) coordinates to the (r, z) coordinates as follows, for a given θ .

$$\begin{aligned} r &= r(\alpha(s), \beta(s)) \\ z &= z(\alpha(s), \beta(s)) \end{aligned} \quad (3)$$

Using (2) and the chain rule we have

$$\lambda^2(r) = [\cos\theta \quad \sin\theta] \begin{bmatrix} r_\alpha^2 + z_\alpha^2 & r_\alpha r_\beta + z_\alpha z_\beta \\ r_\alpha r_\beta + z_\alpha z_\beta & r_\beta^2 + z_\beta^2 \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \quad (4)$$

The principal values of $\lambda^2(\theta)$ are the characteristic values of the matrix in (4). They are

$$\begin{aligned} \lambda_1^2 &= 1/2(A + B + ((A - B)^2 + 4C^2)^{1/2}) \\ \lambda_2^2 &= 1/2(A + B - ((A - B)^2 + 4C^2)^{1/2}) \end{aligned} \quad (5)$$

where

$$\begin{aligned} A &= r_\alpha^2 + r_\beta^2 \\ B &= r_\beta^2 + z_\beta^2 \end{aligned}$$

and

$$C = r_\alpha r_\beta + z_\alpha z_\beta$$

The third stretch ratio, the hoop stretch ratio, is given by

$$\lambda_3^2 = r^2/a^2 \quad (6)$$

TRIANGULAR BILINEAR ELEMENT. The potential energy expression for a deformed elastomer is not quadratic in the variables to be determined. That is, the potential energy minimization problem is nonlinear. In this section we describe a convenient way to compute the gradient and tangent matrices which must be repetitively calculated in the minimization process. We choose a triangular three node element. The node numbering and coordinates for the element are shown in Figure 2. We map the undeformed triangular element to the unit triangle shown in Figure 3 and interpolate (α, β) over the mapped triangle as follows.

$$\begin{aligned}\alpha &= \alpha_1 (1 - \xi - \eta) + \alpha_2 \xi + \alpha_3 \eta \\ \beta &= \beta_1 (1 - \xi - \eta) + \beta_2 \xi + \beta_3 \eta\end{aligned}\tag{7}$$

Using (7) we obtain

$$\begin{bmatrix} d\xi \\ d\eta \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} -d\beta_2 & d\alpha_2 \\ -d\beta_3 & d\alpha_3 \end{bmatrix} \begin{bmatrix} d\alpha \\ d\beta \end{bmatrix}\tag{8}$$

where

$$\Delta = d\alpha_2 d\beta_3 - d\alpha_3 d\beta_2$$

and

$$\begin{aligned}d\alpha_1 &= \alpha_3 - \alpha_2 & d\alpha_2 &= \alpha_1 - \alpha_3 & d\alpha_1 + d\alpha_2 + d\alpha_3 &= 0 \\ d\beta_1 &= \beta_3 - \beta_2 & d\beta_2 &= \beta_1 - \beta_3 & d\beta_1 + d\beta_2 + d\beta_3 &= 0\end{aligned}$$

Similarly we interpolate (r, z)

$$\begin{aligned}r &= r_1 (1 - \xi - \eta) + r_2 \xi + r_3 \eta \\ z &= z_1 (1 - \xi - \eta) + z_2 \xi + z_3 \eta\end{aligned}\tag{9}$$

from which

$$\begin{bmatrix} dr \\ dz \end{bmatrix} = \begin{bmatrix} dr_3 & -dr_2 \\ dz_3 & -dz_2 \end{bmatrix} \begin{bmatrix} d\xi \\ d\eta \end{bmatrix}\tag{10}$$

Using (8) and (10) we can obtain r_α , r_β , z_α , and z_β in terms of the nodal variables $(r_1, z_1, r_2, z_2, r_3, z_3)$ and the undeformed geometry. We find

$$r_{\alpha} = - \frac{1}{\Delta} \sum r_i d\beta_i \quad r_{\beta} = - \frac{1}{\Delta} \sum r_i d\beta_i \quad i = 1, 2, 3 \quad (11)$$

$$z_{\alpha} = - \frac{1}{\Delta} \sum z_i d\beta_i \quad z_{\beta} = - \frac{1}{\Delta} \sum z_i d\alpha_i$$

We note the relations in (11) can be used in (5) and (6) to compute the principal stretch ratios λ_i ($i=1,2,3$). The quantities r_{α} , z_{α} , r_{β} , z_{β} in (11) are constant in an element. It is useful to express these quantities as vector dot products since they are needed to calculate the potential energy and thus will be involved in the computation of the gradient and tangent matrices. We define

$$\begin{aligned} p^T &= -1/\Delta [d\beta_1 \ 0 \ d\beta_2 \ 0 \ d\beta_3 \ 0] \\ q^T &= -1/\Delta [0 \ d\beta_1 \ 0 \ d\beta_2 \ 0 \ d\beta_3] \\ r^T &= 1/\Delta [d\alpha_1 \ 0 \ d\alpha_2 \ 0 \ d\alpha_3 \ 0] \\ s^T &= 1/\Delta [0 \ d\alpha_1 \ 0 \ d\alpha_2 \ 0 \ d\alpha_3] \end{aligned} \quad (12)$$

then

$$\begin{aligned} r_{\alpha} &= u^T p & z_{\alpha} &= u^T s \\ r_{\beta} &= u^T r & z_{\beta} &= u^T s \end{aligned} \quad (13)$$

where

$$u^T = (r_1, z_1, r_2, z_2, r_3, z_3)$$

In addition, we will need a similar expression for the radius used to calculate λ_3 in (6). Since the geometry (i.e. r and z) are interpolated with bilinear functions we will attempt to use one point integration to compute the expressions in the energy. The form of r at the center of the element, r_c , becomes

$$r_c = (r_1 + r_2 + r_3)/3 = u^T t \quad (14)$$

where

$$t^T = 1/3 [1 \ 0 \ 1 \ 0 \ 1 \ 0]$$

We now outline the computation of the gradient and tangent matrices of the potential energy. From the Valanis and Landel form [6] of the internal energy we have

$$\Pi = \int_{\tau} U(\lambda_1, \lambda_2, \lambda_3) d\tau - W \quad (15)$$

where

$$U(\lambda_1, \lambda_2, \lambda_3) = \sum F(\lambda_i)$$

τ = the volume of an element

W = the work done by external forces

Using one point integration we have

$$\Pi = \pi r_c \Delta U - W \quad (16)$$

Then the gradient and tangent matrices become

$$g = \frac{\partial \Pi}{\partial u^T} = \pi r_c \Delta \frac{\partial U}{\partial u^T} - \frac{\partial W}{\partial u^T} \quad (17)$$

and

$$k = \frac{\partial^2 \Pi}{\partial u \partial u^T} = \pi r_c \Delta \frac{\partial^2 U}{\partial u \partial u^T} - \frac{\partial^2 W}{\partial u \partial u^T} \quad (18)$$

Since $\lambda_i = \lambda_i(r_1, z_1, r_2, z_2, r_3, z_3) = \lambda_i(u)$ for an element we can then write

$$U = U(u) \quad (19)$$

and directly compute the gradient of U as follows.

$$\frac{\partial U}{\partial u^T} = \sum_{i=1}^3 U_i \lambda_i', u \quad (20)$$

where

$$U_i = U, \lambda_i$$

and

$$\lambda_{i,u} = \begin{bmatrix} \lambda_{i,r1} \\ \lambda_{i,z1} \\ \lambda_{i,z3} \end{bmatrix}$$

Similarly, the tangent matrix of U is determined from

$$\begin{aligned} \frac{\partial^2 U}{\partial u \partial u}^T = & \sum_{i=1}^3 [U_{ii} \lambda_{i,u} \lambda_{i,u}^T + U_i \lambda_{i,uu}] \\ & + \sum_{\substack{i=1 \\ j=2,3 \ (j>i)}} U_{ij} [\lambda_{i,u} \lambda_{j,u}^T + \lambda_{j,u} \lambda_{i,u}^T] \end{aligned} \quad (21)$$

where

$$\lambda_{i,uu} = \begin{bmatrix} \lambda_{i,r1r1} & \lambda_{i,r1z1} & \lambda_{i,r1r2} & \cdots \\ & \lambda_{i,z1z1} & \lambda_{i,z1z2} & \cdots \\ \text{(sym)} & & \lambda_{i,r2r2} & \cdots \end{bmatrix}$$

and

$$\lambda_{i,u} \lambda_{j,u}^T = \begin{bmatrix} \lambda_{i,r1} \\ \lambda_{i,z1} \\ \lambda_{i,z3} \end{bmatrix} \begin{bmatrix} \lambda_{j,r1} & \lambda_{j,z1} & \cdots & \lambda_{j,z3} \end{bmatrix}$$

Note, $\lambda_{3,uu} = 0$ so there are eight coefficient matrices in (21). Using (5), (6) and (14) we have

$$\lambda_{i,u} = \lambda_{i,A}A_{,u} + \lambda_{i,B}B_{,u} + \lambda_{i,C}C_{,u} \quad i = 1,2 \quad (22)$$

and

$$\lambda_{3,u} = t/\alpha$$

Similarly,

$$\begin{aligned} \lambda_{i,uu} = & \lambda_{i,A}A_{,uu} + \lambda_{i,B}B_{,uu} + \lambda_{i,C}C_{,uu} + \lambda_{i,AA}A_{,u}A_{,u}^T + \\ & \lambda_{i,BB}B_{,u}B_{,u}^T + \lambda_{i,CC}C_{,u}C_{,u}^T + \lambda_{i,AB}(A_{,u}B_{,u}^T + B_{,u}A_{,u}^T) \\ & + \lambda_{i,AC}(A_{,u}C_{,u}^T + C_{,u}A_{,u}^T) + \lambda_{i,BC}(B_{,u}C_{,u}^T + C_{,u}B_{,u}^T) \\ i = & 1,2 \end{aligned} \quad (23)$$

and

$$\lambda_{3,uu} = 0$$

The expressions in (22) and (23) can be conveniently computed using the following relations obtained from definitions above.

$$\begin{aligned} A &= u^T[pp^T + qq^T]u \\ B &= u^T[rr^T + ss^T]u \\ C &= u^T[pr^T + rp^T]u \end{aligned} \quad (24)$$

Then,

$$\begin{aligned} A_{,u} &= 2[r_\alpha p + z_\alpha q] & B_{,u} &= 2[r_\beta r + z_\beta s] \\ A_{,uu} &= 2[pp^T + qq^T] & B_{,uu} &= 2[rr^T + ss^T] \end{aligned} \quad (25)$$

and

$$\begin{aligned} C_{,u} &= r_\alpha r + r_\beta p + z_\alpha s + z_\beta q \\ C_{,uu} &= pr^T + rp^T + qs^T + sq^T \end{aligned} \quad (26)$$

The only remaining tasks are to determine the derivatives of the stretch ratios and to perform the axisymmetric integration of the element gradient and tangent matrices. Using (5) we have

$$\begin{aligned} \lambda_{1,A} &= \frac{1}{4\lambda_1} \left[1 + \frac{A-B}{\lambda_1^2 - \lambda_2^2} \right] \\ \lambda_{1,B} &= \frac{1}{\lambda_1} \left[\frac{1}{2} - \lambda_1 \lambda_{1,A} \right] \end{aligned}$$

$$\lambda_{1,C} = \frac{C}{\lambda_1 [\lambda_1^2 - \lambda_2^2]}$$

$$\lambda_{2,A} = \frac{1}{\lambda_2} \left[\frac{1}{2} - \lambda_1 \lambda_{1,A} \right] \quad (27)$$

$$\lambda_{2,B} = \frac{\lambda_1 \lambda_{1,A}}{\lambda_2}$$

$$\lambda_{2,C} = \frac{-\lambda_1 \lambda_{1,C}}{\lambda_2}$$

$$\lambda_{1,AA} = \frac{1}{\lambda_1} \left[\frac{1}{4} \frac{(\lambda_1^2 - \lambda_2^2) - 2(A - B)(\lambda_1 \lambda_{1,A} - \lambda_2 \lambda_{2,A})}{(\lambda_1^2 - \lambda_2^2)^2} - \lambda_{1,A}^2 \right]$$

$$\lambda_{1,AB} = -\lambda_{1,AA} - \frac{\lambda_{1,A}}{2\lambda_1}$$

$$\lambda_{1,AC} = \frac{1}{\lambda_1} \left[\frac{2C(\lambda_2 \lambda_{2,A} - \lambda_1 \lambda_{1,A})}{(\lambda_1^2 - \lambda_1^2)^2} - \lambda_{1,A} \lambda_{1,C} \right]$$

$$\lambda_{2,AA} = -\frac{1}{\lambda_2} [\lambda_{1,A}^2 + \lambda_{2,A}^2 + \lambda_1 \lambda_{1,AA}]$$

$$\lambda_{2,AB} = -\frac{1}{\lambda_2} [\lambda_{1,A} \lambda_{1,B} + \lambda_{2,A} \lambda_{2,B} + \lambda_1 \lambda_{1,AB}]$$

$$\lambda_{2,AC} = -\frac{1}{\lambda_2} [\lambda_{1,A} \lambda_{1,C} + \lambda_{2,A} \lambda_{2,C} + \lambda_1 \lambda_{1,AC}]$$

$$\lambda_{1,BB} = -\frac{1}{\lambda_1} [\lambda_{1,A} \lambda_{1,B} + \lambda_{1,B}^2 + \lambda_1 \lambda_{1,AB}]$$

$$\lambda_{1,BC} = -\frac{1}{\lambda_1} [\lambda_{1,A} \lambda_{1,C} + \lambda_{1,B} \lambda_{1,C} + \lambda_1 \lambda_{1,AC}]$$

$$\lambda_{2,BC} = -\frac{1}{\lambda_2} [\lambda_{1,B} \lambda_{1,C} + \lambda_{2,B} \lambda_{2,C} + \lambda_1 \lambda_{1,BC}]$$

$$\lambda_{1,CC} = \frac{1}{\lambda_1} \left[\frac{(\lambda_1^2 - \lambda_2^2) - 2C(\lambda_1 \lambda_{1,C} - \lambda_2 \lambda_{2,C})}{(\lambda_1^2 - \lambda_2^2)^2} - \lambda_{1,C}^2 \right]$$

and

$$\lambda_{2,CC} = -\frac{1}{\lambda_2} [\lambda_{1,C}^2 + \lambda_{2,C}^2 + \lambda_1 \lambda_{1,CC}]$$

Note, $\lambda_{2,uu}$ is computed using $\lambda_1^2 + \lambda_2^2 = A + B$ so $\lambda_{2,BB}$ is not needed. That is,

$$\lambda_{2,uu} = \frac{1}{\lambda_2} \left[\frac{1}{2} (A_{,uu} + B_{,uu}) - \lambda_{1,u} \lambda_{1,u}^T - \lambda_{2,u} \lambda_{2,u}^T \right] \quad (28)$$

Given any configuration (r,z) we can now compute the approximate gradient and tangent matrices of the potential energy using the above bilinear three noded triangular axisymmetric element. The Newton - Raphson method can then be used to locate extremum values of the potential energy.

MATERIAL MODEL. The internal energy expression used in this effort is valid for a nearly incompressible solid undergoing large deformations [6]. Elastomers fall into this category. The specific form used is

$$U = 1/2 \hat{\lambda} [\ln(\lambda_1 \lambda_2 \lambda_3)]^2 + 2\mu \sum_{i=1}^3 \lambda_i [\ln(\lambda_i) - 1] \quad (29)$$

where $\hat{\lambda}$, μ = the Lamé constants.

Many other forms are available [7]. We utilize the expression in (29) to demonstrate the finite element algorithm. The terms needed in the gradient and tangent matrices, equations (19) and (20), are

$$U_i = \frac{\hat{\lambda}}{\lambda_i} \ln(\lambda_1 \lambda_2 \lambda_3) + 2\mu \ln(\lambda_i) \quad i=1,2,3$$

$$U_{ii} = \frac{1}{\lambda_i} [\hat{\lambda} [1 - \ln(\lambda_1 \lambda_2 \lambda_3)] + 2\mu \lambda_i] \quad i=1,2,3 \quad (30)$$

and

$$U_{ij} = \frac{\hat{\lambda}}{\lambda_i \lambda_j} \quad i=1,2 \quad j=2,3 \quad j>i$$

END THRUST AND PROBE PENETRATION. To demonstrate the above algorithm we solved two problems. The first is the "end thrust" problem in which a cylinder is compressed by first bonding its ends to plates. The cylinder and plates are then placed in a hydraulic tensile/compression tester which forces the plates towards each other. The second is the "probe penetration" problem which involves forcing a probe with a hemispherical end down the axis of the cylinder.

A. END THRUST. A cylinder of radius 1.0 in and height 2.0 in was assumed. The Lamé constants chosen were $\hat{\lambda} = 16,000$ psi and $\mu = 160$ psi. These values closely represent styrene - butadiene rubber. Solutions were obtained by enforcing motion of the top surface of the cylinder. Results are shown in Figure 4 for a 10% reduction in height. The deformed mesh and profiles of λ_1, λ_2 and λ_3 are shown.

B. PROBE PENETRATION. For this example a cylinder of radius 1.0 in and height 1.0 in were assumed. The values of $\hat{\lambda}$ and μ were the same as those used for the end thrust problem. Some additional comments relating to the method used to represent contact are worthwhile. A penalty method, in which the distance between nodes (which have moved "inside" the probe surface) and the probe surface is minimized, was used to model contact. One of two terms was added to the potential energy, depending on the location of the node. The first term (see Figure 4.) is

$$\Pi_n' = \gamma_{np} [r_p^2 - (r_n^2 + (z_n - c)^2)^{1/2}]^2 \quad (31)$$

where

$$\gamma_{np} = p k_{znzn} ,$$

p = a user defined constant to weight the penalty term,

k_{znzn} = the diagonal term of the tangent matrix associated with a vertical displacement,

(r_n, z_n) = the coordinates of the node,

c = the height to the center of the hemispherical end of the probe,

and

r_p = the radius of the hemispherical end of the probe.

The second term is

$$\Pi_n'' = \gamma_{np} [r_p - r_n]^2 \quad (32)$$

The Π_n' term is used if the node is inside the spherical end of the probe and the Π_n'' term for the case when the node is inside the cylindrical portion of the probe.

The gradient and tangent matrices were computed for Π_n' and Π_n'' and used to modify the global gradient and tangent matrices when a node was in contact. After a node was determined to be in contact the appropriate penalty term was applied for all succeeding probe locations analyzed. That is, the node was not released. However, none of the nodes in contact moved outside the contact surface for the problems analyzed here. The results for a probe of radius 0.25 in penetrating 0.30 in into the cylinder is shown in Figure 5. Also shown are results obtained using the ABAQUS program with a Mooney - Rivlin material given by

$$U = 80 \text{ psi } [\lambda_1^2 + \lambda_2^2 + \lambda_1^{-2} \lambda_2^{-2} - 3] + 20 \text{ psi } [\lambda_1^{-2} + \lambda_2^{-2} + \lambda_1^2 \lambda_2^2 - 3] \quad (33)$$

The results shown in Figure 5 indicate that there is relatively good agreement between the ABAQUS calculations and the calculations made using the above finite element algorithm. Overall displacement profiles and the location of zero hydrostatic pressure (i.e. approximate zone of near zero volume change) agree well.

CONCLUSION. A finite element algorithm was presented in which large axisymmetric deformations of nearly incompressible materials can be determined. Results were presented for end thrust and probe penetration problems. Different material models can be accommodated in this formulation with relative ease. The computer program written to implement the above finite element algorithm was checked by comparing results obtained using it to results obtained using the ABAQUS finite element code for a probe penetration problem.

ACKNOWLEDGEMENT. The authors would like to thank Cpt. Charles S. White of the Mechanics and Structural Integrity Laboratory at AMMRC for performing the calculations made using the ABAQUS program.

REFERENCES.

- [1] I. Fried, "Nonlinear finite element computation of the equilibrium and stability of the circular plate", Int. J. Num. Meth. Eng. (1981), 1436 - 1440.
- [2] I. Fried, "Finite element computation of large elastic deformations", The mathematics of finite elements and applications IV, MAFELAP (1981), edited by J. R. Whitman, Academic Press, 1982.
- [3] A. R. Johnson, "Large deformations and stability of axisymmetric Mooney membranes - finite element solutions", Trans. of the twenty - eighth conference of Army mathematicians, U.S. Army Research Office Report No. 83 - 1, February 1983.
- [4] A. R. Johnson, "Finite element analysis of fabrics with nonlinear stress - strain laws", Trans. of the first Army conference on applied mathematics and computing, U.S. Army Report No. 84 - 1, February 1984.
- [5] A. R. Johnson and C. J. Quigley, "Buckled elastica in contact - finite element solutions", Trans. of the second Army conference on applied mathematics and computing, U.S. Army Research Office Report No. 85 - 1, February 1985.
- [6] K. C. Valanis and R. F. Landel, "The strain - energy function of a hyperelastic material in terms of the extension ratios", J. of Applied Physics, Vol. 38, No. 7, June 1967, 2997 - 3002.
- [7] L. R. G. Treloar, "The mechanics of rubber elasticity", Proc. R. Soc. Lond. A. 351, 1976, 301 - 330.

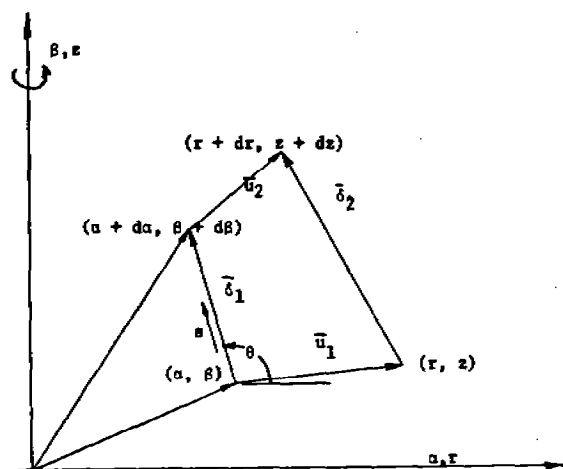


Figure 1. Notation for material line deformations.

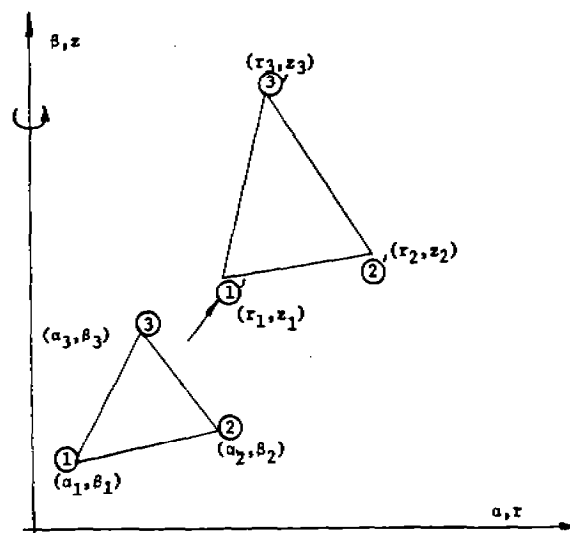


Figure 2. Node numbering and coordinates for element.

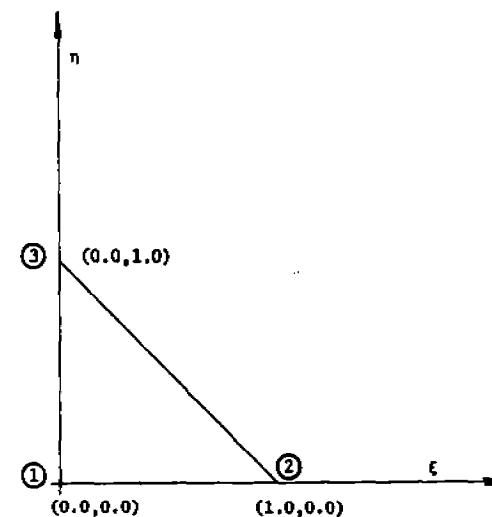


Figure 3. Unit triangle used for interpolation.

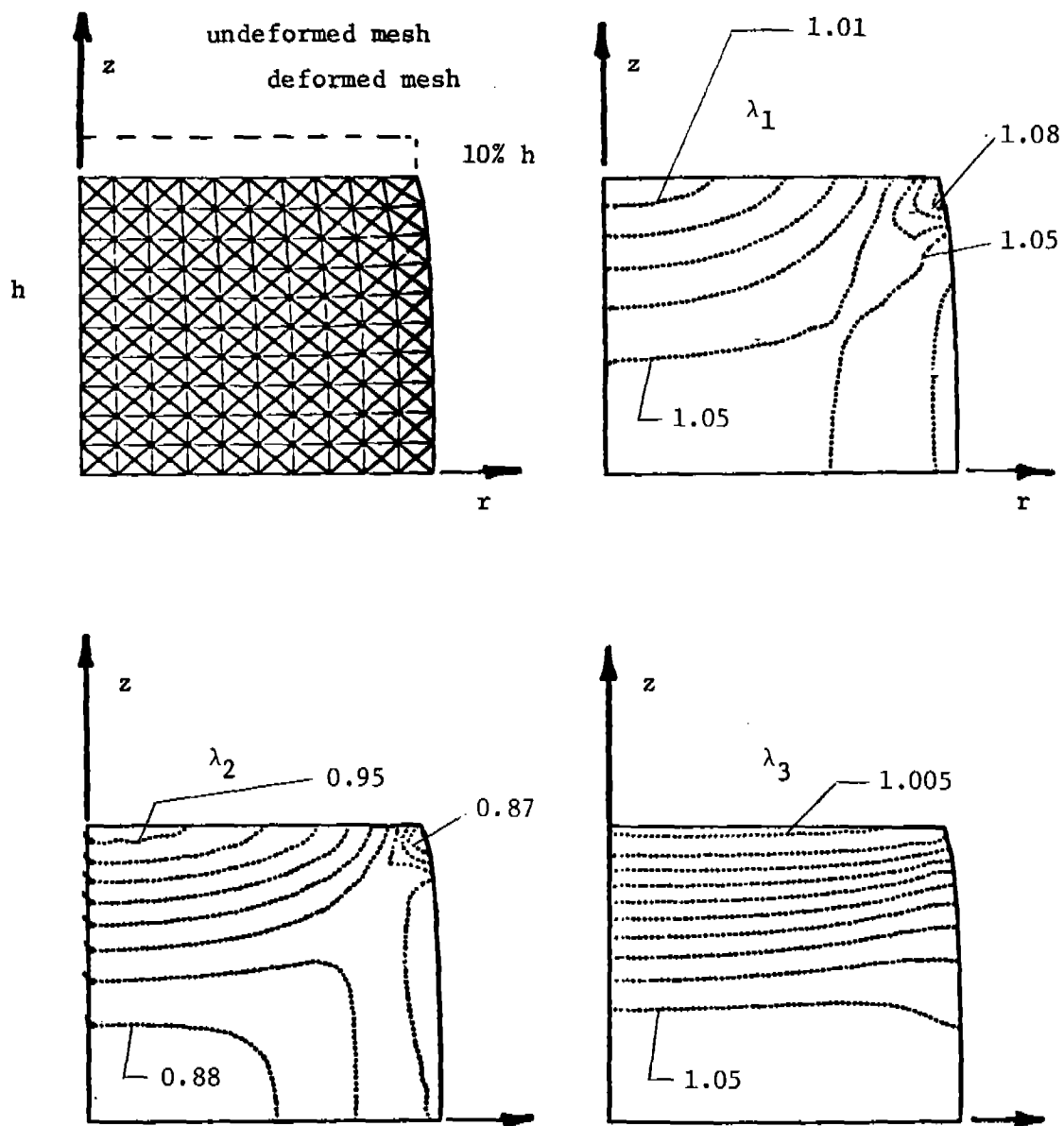


Figure 4. End Thrust: Deformed mesh at 10% reduction and profiles of stretch ratios.

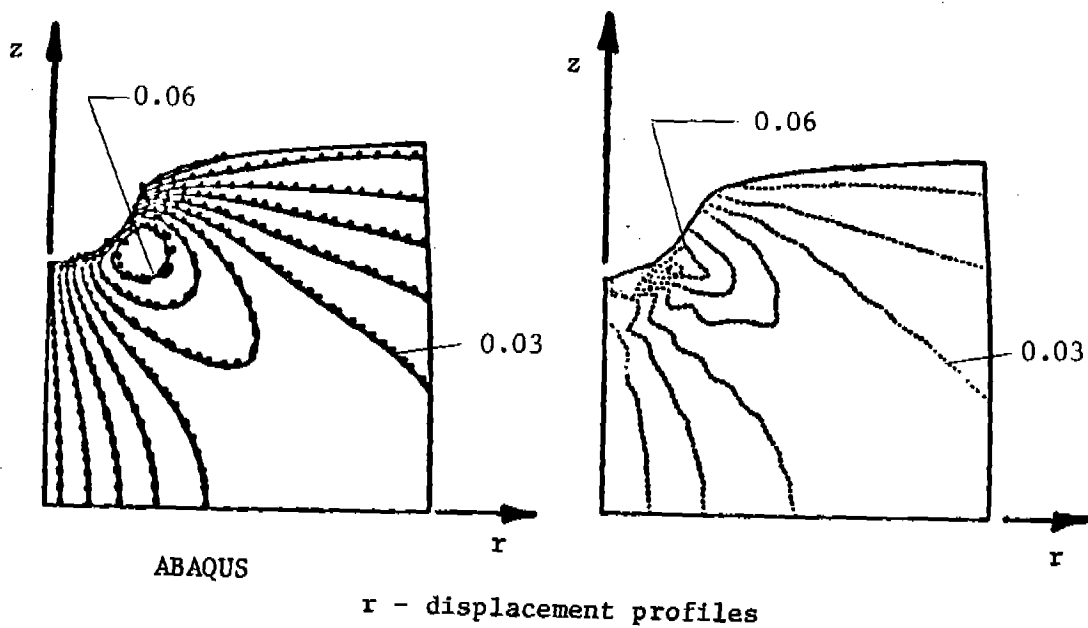
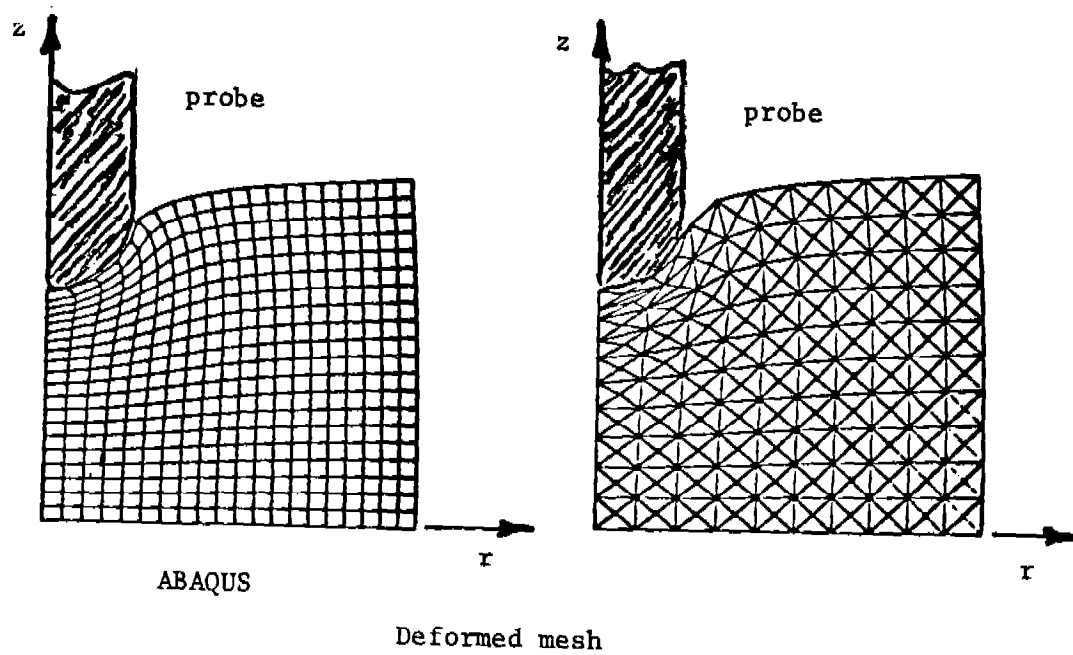
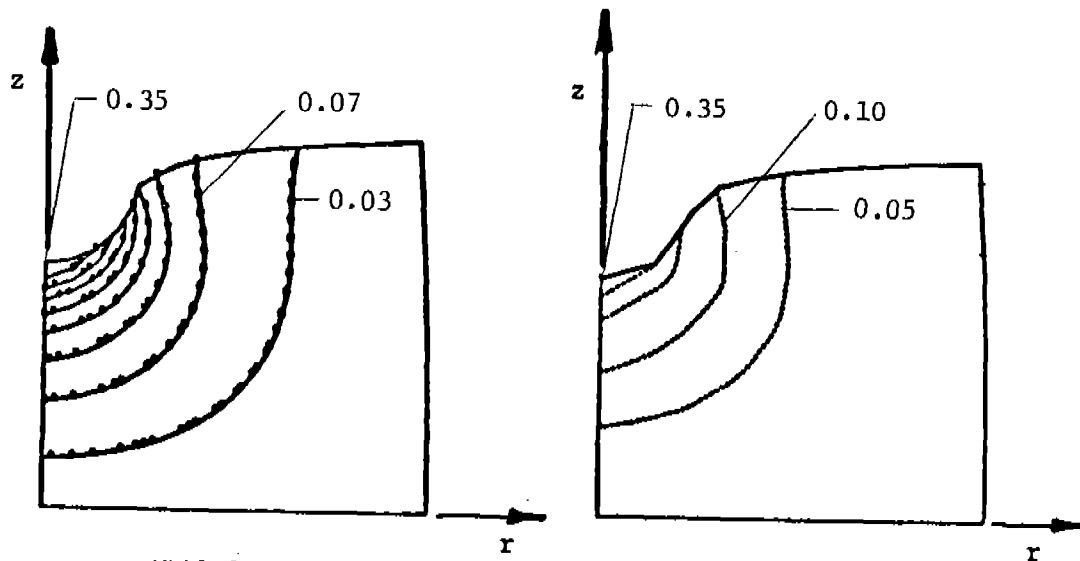
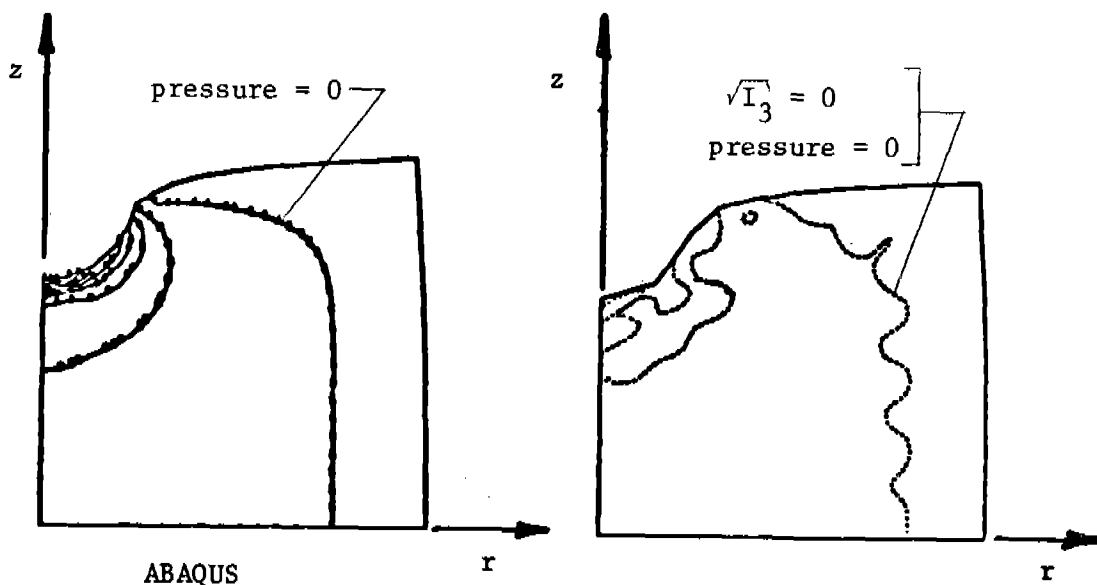


Figure 5. Probe Penetration: Deformed mesh at 35% probe penetration, profiles for r, z displacements and hydrostatic pressure. (continued next page)



z - displacement profiles



hydrostatic pressure profiles

$\sqrt{I_3}$ - profiles

Figure 5. Continued.

A TECHNIQUE FOR CALCULATING PATH INTEGRALS
FOR NONLINEAR FRACTURE

J.R. Whiteman[†] and G.M. Thompson^{*}
Institute of Computational Mathematics
Department of Mathematics and Statistics
Brunel University, Uxbridge, Middlesex, UB8 3PH, England

I. INTRODUCTION. Path independent integrals for fracture mechanics were presented by Eshelby [5] and Rice [7], and a number of papers extending the range of application, e.g. Atluri [1], Blackburn [3], Blackburn et al [4], Hellen [6] and Rice [8], have since appeared.

The present paper describes a finite element method for calculating approximations to a path integral J_p for fracture-problems involving elastic-plastic deformation with hardening. This involves the use of techniques based on incremental plasticity and allows approximations to J_p to be calculated for contours surrounding the crack tip. Recent comments by Atluri and Tracey emphasise the importance of being able to calculate such approximations for contours at varying distances from a crack tip, see Atluri [2] and Tracey [10].

II. J_p -INTEGRAL. Rice [7] proved that, for a homogeneous two-dimensional elastic body, the rate of decrease in potential energy with respect to crack length is equal to the path independent integral, J . That is

$$J = - \frac{\partial P_E}{\partial L}, \quad (2.1)$$

where P_E is the potential energy and L the crack length. J is defined for a crack with flat surfaces parallel to the x_1 -axis, see Fig. 1, as

$$J \equiv \int_{\Gamma} \left[W dx_2 - T_i \frac{\partial u_i}{\partial x_1} ds \right] \quad (2.2)$$

where the contour Γ surrounding the crack tip, starts from the lower crack surface and continues in an anticlockwise direction until it ends on the upper surface, W is the strain energy density, u_i are the displacements, and the tractions T_i are defined with respect to the unit outward normal vector \underline{n} , so that $T_i \equiv \sigma_{ij} n_j$.

[†] Supported in part by the United States Army Research, Development and Standardisation Group, London, England.

^{*} GKN Technological Centre, Design Analysis Group, Birmingham New Road, Wolverhampton, WV4 6BW, England.

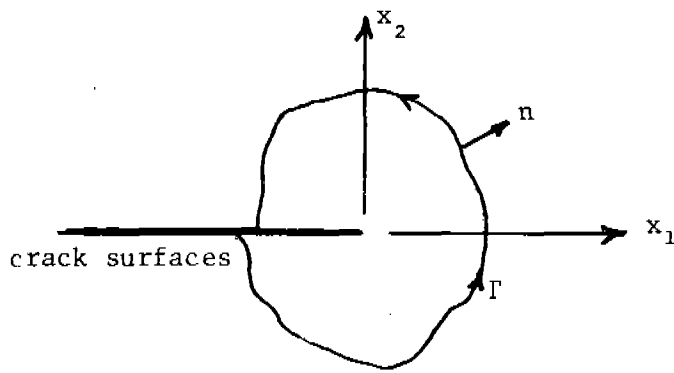


Fig. 1. Two-dimensional crack with contour Γ

For problems of linear elastic fracture the value of J can be used as a fracture criterion. In the present work for elastic-plastic fracture problems a corresponding integral, the J_p -integral, is defined. This integral can be developed from the J -integral, (2.2), by separating what was for the elastic case the strain energy density W into elastic and plastic components, W_e and W_p respectively. Thus

$$W = W_e + W_p, \quad (2.3)$$

where the elastic component is given in terms of the stress and elastic strain components by

$$W_e = \frac{1}{2} \sigma_{ij} (\epsilon_{ij})_e. \quad (2.4)$$

The plastic term, W_p , is defined as

$$W_p \equiv \int_0^{\bar{\epsilon}_p} \bar{\sigma} d\bar{\epsilon}_p \quad (2.5)$$

where $\bar{\sigma}$ and $\bar{\epsilon}_p$ are respectively the effective stress and effective plastic strain. Equations (2.2) - (2.5) together define J_p .

III. CALCULATION OF THE J_p -INTEGRAL

Full details of the method of approximating J_p throughout the load history for elastic-plastic deformation are given by Whiteman and Thompson [11]; a description of the algorithm and use of the MODEL finite element code appears in [9].

The finite element method is applied using a formulation in terms of displacements. With incremental plasticity the load is applied incrementally, and for the k th increment the finite element approximation $du_h^{(k)}$ to the corresponding increment of displacement is first calculated. The associated increments of strain and stress $d\epsilon_h^{(k)}$ and $d\sigma_h^{(k)}$ are then retrieved and the total displacements $u_h^{(k)}$, strains $\epsilon_h^{(k)}$ and stresses $\sigma_h^{(k)}$, for the combined load up to the k th increment, are calculated.

For the approximation of J_p , as in (2.2) - (2.5), to be calculated, a contour path Γ , see Fig. 2, is chosen through a ring of elements Ω^e surrounding the crack tip. The total value of the approximation $(J_p)_h^{(k)}$ is calculated by summing the contributions $(J_p^e)_h^{(k)}$ from elements involving

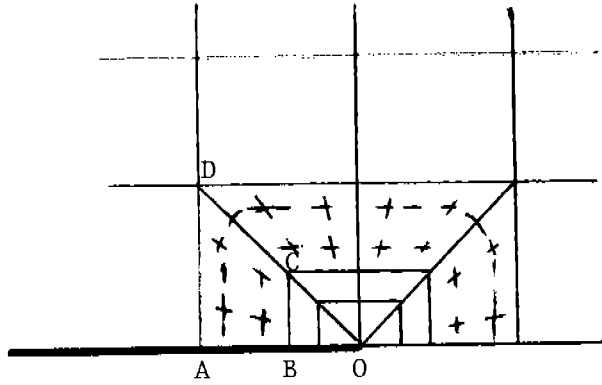


Fig. 2. The contour Γ around the crack tip, point O. The Gauss points are indicated by X and Γ by the dotted line. The upper crack surface is the line OA

segments $\Gamma^{(e)}$ of Γ . Each element Ω^e is transformed in turn onto a *standard element* in the (ξ_1, ξ_2) -plane so that the image of $\Gamma^{(e)}$ is either a line $\xi_1 = \text{Const}$, a line $\xi_2 = \text{Const}$ or a path consisting of parts $\xi_1 = \text{Const}$ and $\xi_2 = \text{Const}$.

Along the image of $\Gamma^{(e)}$, for example $\xi_1 = \text{Const}$, we have that

$$(J_p^e)_h^{(k)} = \sum_j I_3^{(e)} \left(\xi_1^{(i)}, \xi_2^{(j)} \right)_h^{(k)} G_j, \quad (3.1)$$

where the $(\xi_1^{(i)}, \xi_2^{(j)})$ are Gauss points and the G_j are the corresponding weights. If the stresses and displacements at the end of the k^{th} load increment are respectively $(\sigma_{ij}^e)_h^{(k)}$ and $(u_i^e)_h^{(k)}$, $i, j = 1, 2$, then $I_3^{(e)}(\dots)_h^{(k)}$ is given by

$$\begin{aligned} I_3^{(e)}(\dots)_h^{(k)} &\equiv \frac{1}{2} \left[(\sigma_{11}^e)_h^{(k)} \frac{\partial (u_1^e)_h^{(k)}}{\partial x_1} + (\sigma_{12}^e)_h^{(k)} \left(\frac{\partial (u_1^e)_h^{(k)}}{\partial x_2} + \frac{\partial (u_2^e)_h^{(k)}}{\partial x_1} \right) \right. \\ &\quad \left. + (\sigma_{22}^e)_h^{(k)} \frac{\partial (u_2^e)_h^{(k)}}{\partial x_2} + W_p^{(k)} \right] \frac{\partial x_2}{\partial \xi_2} \\ &- \left[\left\{ (\sigma_{11}^e)_h^{(k)} n_1 + (\sigma_{12}^e)_h^{(k)} n_2 \right\} \frac{\partial (u_1^e)_h^{(k)}}{\partial x_1} + \left\{ (\sigma_{12}^e)_h^{(k)} n_1 + (\sigma_{22}^e)_h^{(k)} n_2 \right\} \right. \\ &\quad \left. \frac{\partial (u_2^e)_h^{(k)}}{\partial x_1} \right] \left[\left(\frac{\partial x_1}{\partial \xi_2} \right)^2 + \left(\frac{\partial x_2}{\partial \xi_2} \right)^2 \right]^{\frac{1}{2}}, \quad (3.2) \end{aligned}$$

where $W_p^{(k)}$ is the plastic work term at the end of the k^{th} increment

$$\text{given by } (W_p^{(k)}) = \int_0^{\bar{\epsilon}_p^{(k)}} \bar{\sigma}^{(k)} d\bar{\epsilon}_p^{(k)}.$$

For nonlinear fracture problems, where incremental techniques are used to model the elastic-plastic deformation, if the total load is applied through L increments, then approximations to $(J_p)_h^{(k)}$ can be calculated for each load increment $k=1,2,\dots,L$. In each increment a number of approximations are calculated for different contours Γ which are at different distances from the crack tip O , see Fig. 2. These approximations, together with their mean value, are used as criteria for fracture.

Approximations to J_p have been calculated for a number of contours Γ as the load is incremented for several two-dimensional Mode I crack problems. In each case a power law relating effective plastic strain and effective stress has been employed. Details of the problems together with some numerical results are given by Thompson and Whiteman in [9] and [11].

REFERENCES

1. Atluri, S.N., Path-independent integrals in finite elasticity and inelasticity, with body forces, inertia and arbitrary crack face conditions. Eng. Fracture Mech. 16, 341-364, 1982.
2. Atluri, S.N., Computational aspects of finite strain inelastic solid and fracture mechanics. Proc. Third Army Conf. on Applied Mathematics and Computing, Atlanta, May 1985.
3. Blackburn, W.S., Path independent integrals to predict onset of crack instability in an elastic plastic material. Int. J. Fracture 8, 343-356, 1972.
4. Blackburn, W.S., Hellen, T.K. and Jackson, A.D., An integral associated with the state of a crack tip in a non-elastic material. Int. J. Fracture 13, 183-200, 1977.
5. Eshelby, J.D., The continuum theory of lattice defects. pp.79-144, Solid State Physics, Vol.III. Academic Press, New York, 1956.
6. Hellen, T.K., Numerical methods in fracture mechanics. pp.145-181 of G.C. Chell (ed.), Developments in Fracture Mechanics - 1. Applied Science, London, 1981.
7. Rice, J.R., A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. 35, 379-386, 1968.
8. Rice, J.R., Elastic-plastic fracture mechanics. pp.23-54 of F. Erdogan (ed.), The Mechanics of Fracture. ASME-AMD 19, ASME, New York, 1976.
9. Thompson, G.M. and Whiteman, J.R., The use of the MODEL finite element code in the solution of problems of linear elastic and nonlinear fracture. Technical Report BICOM 84/4. Institute of Computational Mathematics, Brunel University, 1984.
10. Tracey, D.M. and Freese, C.M., Crack solutions and ductile fracture criteria. Proc. Third Army Conf. on Applied Mathematics and Computing. Atlanta, May 1985.
11. Whiteman, J.R. and Thompson, G.M., Finite element calculations of parameters for singularities in problems of fracture. pp.27-47 of J.R. Whiteman (ed.), The Mathematics of Finite Elements and Applications V, MAFELAP 1984. Academic Press, London, 1985.

AN EXPLICIT A PRIORI ASSESSMENT OF SHEAR LOCKING
IN A TRIANGULAR MINDLIN-TYPE PLATE ELEMENT

Alexander Tessler

Mechanics of Materials Branch
Army Materials and Mechanics Research Center
Watertown, Massachusetts 02172

ABSTRACT. An a priori, explicit algebraic procedure for identifying shear locking and excessive solution stiffening in thin shear-deformable plates is explored. Only element level solutions of element Kirchhoff modes are required to establish the nodal degree-of-freedom constraints. These constraints clearly identify whatever kinematic stiffening might exist. The methodology is demonstrated by the use of a conforming, three-node Mindlin element. Several discretizations of square plates are examined. The results are compared, and fully confirmed, with the corresponding numerical solutions. Examples of alternate discretizations, which alleviate the locking effect, are also presented.

1. INTRODUCTION. The locking phenomenon, intrinsic to C^0 penalty constraint models, is by far the most severe detriment to element performance. Encountered in the limiting penalty regime (viz., when the penalty parameter approaches infinity while the penalty strains diminish to zero), locking is evidenced by grossly erroneous results. Various remedial techniques have been employed in an effort to abate 'shear locking' in shear-deformable flexure elements and mathematically similar 'incompressibility locking' in incompressible elasticity, plasticity and fluid flow formulations. Notable improvements have been achieved by discrete penalty constraints [1-5], reduced integration procedures [6-22], improved penalty strain interpolations [23-25], and penalty parameter modifications [13,18,26-30].

Despite the locking-free characteristics generally achieved by these enhancing schemes, evidence of locking under certain discretizational conditions still remains (e.g., refer to [17,29,30]). Presently, there exists a lack of insight with regard to the actual mechanism of locking, in general, and in these, occasionally locking models, in particular. Attempts have been made to predict locking on the basis of a constraint index criterion [10] and, recently, by operational procedures involving consistency comparisons of discrete and exact equations of equilibrium/motion [31,32]. While the former approach is only a heuristic measure of locking, the latter methods, despite their theoretical merit, appear too cumbersome for practical applications.

The objective of our present effort is to explore in detail the mechanisms of shear locking by, what appears to be, the most direct and effective way for a discretization assessment conducted a priori to the full-scale finite element computation. Herein, we propose an explicit, element-by-element analysis of element Kirchhoff modes (i.e., vanishing coefficients of shear strain polynomial terms) which are responsible for enforcing a nearly shearless element deformation state. Such an analysis involves simultaneously solving a small set of linear algebraic constraint equations for each element, from which kinematic constraints upon individual degrees of freedom (dof) are determined. By starting with elements most severely restrained by boundary conditions and marching across the mesh from element to element, one can definitively identify the true origin of locking within an element and the conditions of locking propagation throughout the whole discretizational domain. In the absence of shear locking, clear evidence of a global (mesh-wide) stiffening effect, if such exists, can be established.

This a priori analysis will be demonstrated on several thin plate problems modeled by a conforming, three-node (nine dof) Mindlin element [29] (refer to Figs.1 and 2 for the plate and element notations). The element, herein referred to as MIN3_{*}, utilizes 'anisoparametric' kinematic interpolations and exact quadratures of all energy and work terms in the variational statement. It also uses the standard Mindlin plate shear correction factor, k_*^2 ($k_*^2 = \pi^2/12$ for isotropic, homogeneous plates [33]) and, hence, constitutes a true penalty constraint element for which the penalty parameter becomes infinitely large as its thickness diminishes to zero. Essentially the same element, termed MIN3 [29], with the shear stiffness (penalty parameter) enhanced by the so-called element-appropriate shear correction factor, $k^2 = (k_* \phi)^2$, is completely devoid of locking (ϕ^2 is an element shear correction dependent upon element material and geometric properties). Yet, it should not be regarded as a true penalty element owing to the finite-valued upper bound of its 'penalty' parameter (for details of this concept and its implementation, refer to [28-29]). For comparison purposes, MIN3's results will also be presented.

In section 2, from the kinematic variable assumptions we derive explicit polynomial expressions for MIN3_{*}'s transverse shear strains and bending curvatures, and define two distinct types of Kirchhoff modes that control element performance.

In section 3, for the purpose of demonstrating the typical MIN3_{*} behavior with its occasional shear locking characteristics, numerical results are presented for several square plate problems discretized by three standard mesh patterns.

In section 4, the proposed explicit procedure for identifying shear locking and related excessive plate stiffening is

elaborated on using standard discretizational patterns. Closed form solutions, clearly exhibiting the mechanisms of mesh constraining, are carried out for the basic macro-element and refined multi-element discretizations. These a priori predictions are confirmed throughout by the numerical solutions of section 3. The main shortcomings of some of the considered meshes are overcome by alternative discretizations along plate boundaries (section 5). Conclusions and some numerical results with our trouble-free MIN3 element are given in section 6.

2. KIRCHHOFF CONSTRAINTS. The transverse shear strains are derived in a consistent manner from the assumptions of the transverse displacement, w , and normal rotations, θ_x and θ_y (the interested reader should consult [29] for details of the derivation of w), which may be written as

$$w = \underline{\zeta} \underline{w} + \underline{L} \frac{\theta_x}{x} + \underline{M} \frac{\theta_y}{y}, \quad (1a)$$

$$\theta_x = \underline{\zeta} \frac{\theta_x}{x}, \quad \theta_y = \underline{\zeta} \frac{\theta_y}{y} \quad (1b)$$

where

$$\underline{w}^T = \{w_i\}, \quad \underline{\theta_x}^T = \{\theta_{xi}\}, \quad \underline{\theta_y}^T = \{\theta_{yi}\} \quad (i = 1, 2, 3) \quad (2)$$

are vectors of nodal dof; the shape functions

$$\underline{\zeta} = \{\zeta_i\}, \quad \underline{L} = \{L_i\}, \quad \underline{M} = \{M_i\} \quad (i = 1, 2, 3) \quad (3)$$

are given in terms of area-parametric coordinates as

$$L_i = 1/8 (b_k N_{i+3} - b_j N_{k+3}),$$

$$M_i = 1/8 (a_j N_{k+3} - a_k N_{i+3}),$$

$$N_{i+3} = 4\zeta_i \zeta_j$$

$$\zeta_i = (c_i + b_i x + a_i y)/2A, \quad \zeta_i = 1 \text{ (summation on } i)$$

(A is the area of a triangle)

$$a_i = x_k - x_j, \quad b_i = y_j - y_k, \quad c_i = x_j y_k - x_k y_j$$

$$(i = 1, 2, 3; \quad j = 2, 3, 1; \quad k = 3, 1, 2). \quad (4)$$

The transverse shear strains are computed using (1) from the relations

$$\gamma_{xz} = w',_x + \theta_y \quad (5a)$$

$$\gamma_{yz} = w',_y + \theta_x \quad (5b)$$

After performing the straightforward algebraic manipulations, we arrive at the following expressions for the shear strains

corresponding to an arbitrary-shape triangle:

$$\gamma_{xz} = A_0 + A_1 y \quad (6)$$

where

$$A_0 = \underline{t}_a \underline{w} + \underline{l}_a \underline{\theta}_x + \underline{m}_a \underline{\theta}_y \quad (7a)$$

$$A_1 = \underline{p}_a \underline{\theta}_x + \underline{q}_a \underline{\theta}_y \quad (7b)$$

and components of the 1×3 row vectors \underline{t}_a , \underline{l}_a , \underline{m}_a , \underline{p}_a and \underline{q}_a are given as

$$\begin{aligned} t_{ai} &= b_i/2A, \\ l_{ai} &= b_i(b_k c_j - b_j c_k)/8A^2, \\ m_{ai} &= [c_i + b_i(a_j c_k - a_k c_j)/2A]/4A, \\ p_{ai} &= -b_i/4A, \\ q_{ai} &= a_i/4A. \\ (i &= 1,2,3; \quad j = 2,3,1; \quad k = 3,1,2). \end{aligned} \quad (8)$$

Similarly, γ_{yz} may be expressed as

$$\gamma_{yz} = B_0 + B_1 x. \quad (9)$$

with

$$B_0 = \underline{t}_b \underline{w} + \underline{l}_b \underline{\theta}_x + \underline{m}_b \underline{\theta}_y, \quad (10a)$$

$$B_1 = \underline{p}_b \underline{\theta}_x + \underline{q}_b \underline{\theta}_y. \quad (10b)$$

where components of the 1×3 row vectors \underline{t}_b , \underline{l}_b , \underline{m}_b , \underline{p}_b and \underline{q}_b are given by the geometric relations

$$\begin{aligned} t_{bi} &= a_i/2A \\ l_{bi} &= [c_i + a_i(c_j b_k - b_j c_k)/2A]/4A \\ m_{bi} &= a_i(a_j c_k - a_k c_j)/8A^2 \\ p_{bi} &= b_i/4A \\ q_{bi} &= -a_i/4A \\ (i &= 1,2,3; \quad j = 2,3,1; \quad k = 3,1,2). \end{aligned} \quad (11)$$

REMARK 2.1 Interestingly, the shear strain approximations (6) and (9) satisfy the following three differential equations:

$$\gamma_{xz}'_x = 0, \quad \gamma_{yz}'_y = 0, \quad \gamma_{xz}'_y + \gamma_{yz}'_x = 0 \quad (12)$$

This means that the edge constraint procedure, used to condense out the mid-edge w dof [29], is equivalent to enforcing differential constraints (12) across the entire element domain. Further, the third equation in (12) may be recognized as an exact transverse shear equilibrium statement for a plate without interior loads. This, of course, is always the case for MIN3*, where any applied load is transformed into consistent nodal forces. Thus, our 'displacement' element also satisfies one plate equilibrium equation exactly.

In the thin (classical) plate regime as the plate span-to-thickness ratio (L/h) approaches infinity, the Kirchhoff constraints

$$\gamma_{xz} \rightarrow 0 \text{ and } \gamma_{yz} \rightarrow 0 \quad (13a)$$

are enforced over the entire element domain assuming that $A/h^2 \rightarrow \infty$ as well (i.e., the element penalty parameter, which is proportional to A/h^2 , must be large). Owing to (6,9), the Kirchhoff modes must be enforced simultaneously, i.e.,

$$A_0 \rightarrow 0, \quad B_0 \rightarrow 0, \quad A_1 \rightarrow 0 \quad (\text{note } A_1 = -B_1) \quad (13b)$$

The single important feature of constraints (13b) is that in their unrestrained form (i.e., prior to applying boundary restraints) each mode contains contributions of at least two independent kinematic variables (i.e., A_0 and B_0 are functions of w , θ_x and θ_y ; A_1 depends on θ_x and θ_y). Such modes will be referred to as the 'true' Kirchhoff (or penalty) modes. Upon their enforcement, a dependence of at least two originally independent fields is achieved. On the other hand, we shall term a Kirchhoff mode 'spurious' if it is only in terms of a single kinematic variable. This latter situation is inherent in isoparametric formulations (i.e., where w , θ_x and θ_y are interpolated by the same order polynomials; refer to [29] for further discussion on the subject), and it is also possible with MIN3* for some geometries and boundary restraints (due to the removal of certain dof from A_i , B_i coefficients).

In order to ascertain the Kirchhoff constraint implications upon the bending strain energy, it will prove insightful to examine the element normal and twisting curvatures:

$$\kappa_x = \theta_{y,x} = b_i \theta_{yi} / 2A, \quad (14a)$$

$$\kappa_y = \theta_{x,y} = a_i \theta_{xi} / 2A, \quad (14b)$$

$$\kappa_{xy} = \theta_{x,x} + \theta_{y,y} = (b_i \theta_{xi} + a_i \theta_{yi}) / 2A. \quad (14c)$$

(summation on i is implied, $i=1,2,3$)

Note that the element bending strain energy is a function of the

curvatures (14), while the transverse shear energy is in terms of the shear strains (5). The two contributions are superimposed to produce the total element strain energy (for details on the strain energy, stiffness matrices and load vectors for MIN3* consult [29]).

3. STANDARD DISCRETIZATIONS FOR SQUARE PLATES. To ascertain the kind of behavior MIN3* typically exhibits, solutions for uniformly loaded, simply supported and clamped square plate ranging from moderately thick ($L/h=10$) to extremely thin ($L/h=10^6$) are carried out. The A, B and C meshes, depicted in Figs. 3 and 4, are standard 4×4 triangular element subdivisions for a symmetric plate quadrant. The center deflection results, normalized with respect to the appropriate Mindlin theory solutions, are summarized in Table 1.

Note that in the simply supported case (Fig. 3), mesh B tends to lock at $L/h=10^2$, with the solution progressively deteriorating as L/h increases. Meshes A and C, however, yield nonlocking results for this problem, although mesh C solutions are superior. In the clamped case (Fig. 4), both A and B meshes lock for $L/h > 10^2$, while mesh C provides a relatively good solution that is only locally locked.

Apparently, mesh A produces no locking in the simply supported case, but locks under clamped conditions. This indicates the influence of boundary restraints on shear locking. Further, from the comparison of the A and B mesh results in the simply supported case, it becomes evident that locking is also mesh dependent.

In what follows we shall demonstrate that while true Kirchhoff modes yield non-locking solutions, spurious ones generally produce a considerable stiffening effect and, in some cases, full element and/or plate locking.

4. ANALYTIC MESH ASSESSMENTS. The procedure to be undertaken involves simultaneously solving three algebraic constraint equations (13b) for each element, from which the implication on the bending curvatures (14) and bending strain energy can readily be ascertained. By marching through the entire mesh, from element to element, all constraints upon individual dof are solved for without much algebraic difficulty. As the result, we shall be able to pinpoint locking and/or related kinematic stiffening of either local (at some nodes) or global (throughout the mesh) nature.

To assess the suitability of meshes A, B and C for thin plate-bending modeling a priori to the full-scale finite element computations, it is first expedient to determine whether a representative macro-element model (Fig. 5) is capable of a

nonlocking plate response. Then we shall carry out the analysis for larger discretizations starting with elements that are most severely restrained by boundary conditions. In this way, even relatively large discretizations can be examined quickly without computerizing the process. The practical examples are presented below.

4.1 Mesh types A and B. Figure 5a depicts a macro-element model (consisting of two MIN3₊ triangles) for a symmetric quadrant of a square plate. Incorporating element nodal coordinates in (8,11) yields the following Kirchhoff modes and bending curvatures:

element P

$$\begin{aligned} w_2 - w_1 + (a/2)(\theta_{y1} + \theta_{y2}) &= 0 & (A_0 \neq 0), \\ w_3 - w_2 + (a/2)(\theta_{x1} + \theta_{x3} - \theta_{y2} + \theta_{y3}) &= 0 & (B_0 \neq 0), \\ \theta_{x1} - \theta_{x2} - \theta_{y2} + \theta_{y3} &= 0 & (A_1 \neq 0). \end{aligned} \quad (15a)$$

$$\begin{aligned} \kappa_x &= (\theta_{y2} - \theta_{y1})/a, \quad \kappa_y = (\theta_{x3} - \theta_{x2})/a, \\ \kappa_{xy} &= (\theta_{x2} - \theta_{x1} + \theta_{y3} - \theta_{y2})/a. \end{aligned} \quad (15b)$$

element Q

$$\begin{aligned} w_3 - w_4 + a/2(\theta_{x3} - \theta_{x4} + \theta_{y1} + \theta_{y3}) &= 0 & (A_0 \neq 0), \\ w_4 - w_1 + a/2(\theta_{x1} + \theta_{x4}) &= 0 & (B_0 \neq 0), \\ \theta_{x3} - \theta_{x4} + \theta_{y1} - \theta_{y4} &= 0 & (A_1 \neq 0). \end{aligned} \quad (16a)$$

$$\begin{aligned} \kappa_x &= (\theta_{y3} - \theta_{y4})/a, \quad \kappa_y = (\theta_{x4} - \theta_{x1})/a, \\ \kappa_{xy} &= (\theta_{x3} - \theta_{x4} - \theta_{y1} + \theta_{y4})/a. \end{aligned} \quad (16b)$$

A-type mesh: Simply supported edges (1-4) and (3-4) and boundary restraints $w_1=w_3=w_4=\theta_{x1}=\theta_{x2}=\theta_{x4}=\theta_{y2}=\theta_{y3}=\theta_{y4}=0$.

In element Q, all but two dof are fixed, while element P has only three active dof. With these boundary conditions, (15) and (16) become

$$w_2 = -(a/2)\theta_{y1}, \quad w_2 = (a/2)\theta_{x3} \quad (\text{these imply } \theta_{x3} = -\theta_{y1}) \quad (17a)$$

$$\kappa_x = -\theta_{y1}/a, \quad \kappa_y = \theta_{x3}/a, \quad \kappa_{xy} = 0 \quad (17b)$$

$$\theta_{x3} = -\theta_{y1} \quad (18a)$$

$$\kappa_x = \kappa_y = 0, \quad \kappa_{xy} = (\theta_{x3} - \theta_{y1})/a \quad (18b)$$

According to our earlier definition, constraints (17a) and (18a) should be regarded as true Kirchhoff modes since each relates two

independent kinematic fields. The relation (18a) is also theoretically correct due to the intrinsic symmetry condition in this plate problem. Obviously, the bending energies are nonvanishing (because $\kappa_x \neq 0$ and $\kappa_y \neq 0$ in element P, and $\kappa_{xy} \neq 0$ in element Q), and there exists no locking in either of the elements.

Because of the excessive boundary restraints (only three out of twelve dof are not fixed) one would intuitively anticipate excessively stiff results or even locking. Evidently, due to the exclusive enforcement of the true Kirchhoff modes the elements do not lock. On the other hand, the solutions for mesh A

corresponding to $L/h \geq 10^4$ exhibit about 20% error in the maximum deflection (Table 1), which may be regarded as unsatisfactory considering the relatively large number of dof in the discretization. (Note that the exact plate theory solution for this problem requires only few Fourier series terms to capture the bulk of the answer [34].)

To ascertain the source of the 'solution stiffening' for this problem we need only consider (for the sake of avoiding any additional algebra) the linear Kirchhoff modes ($A_1 = -B_1 + 0$). For example, it may be convenient to examine first the R_n and S_n ($n=1,2,\dots,N$) elements that are adjacent to the kinematic symmetry axis x (refer to Fig. 6a). The linear Kirchhoff modes for these elements may be written as

$$\theta_{xi(n+1)} - \theta_{xin} - \theta_{yi(n+1)} + \theta_{yj(n+1)} = 0 \quad (R_n) \quad (19)$$

$$\theta_{xj(n+1)} - \theta_{xjn} - \theta_{yin} + \theta_{yjn} = 0 \quad (S_n) \quad (20)$$

Using the symmetry condition along the x axis (i.e., $\theta_{xin} = 0$ for all n) gives

$$\theta_{yi(n+1)} + \theta_{yj(n+1)} \quad (19a)$$

and employing (19a) in (20) we obtain

$$\theta_{xj(N+1)} + \theta_{xjN} + \theta_{xj(N-1)} + \dots + \theta_{xj2} + \theta_{y11} - \theta_{y1j} = \Delta \neq 0 \quad (20a)$$

The R_n element constraints (19a) are certainly spurious, and they influence the linear Kirchhoff constraints of S_n elements just short of locking: (20a) enforces virtually identical θ_{xjp} dof along j -line, and this obviously produces a severe stiffening effect. In a similar fashion we find that the adjacent row of elements is subject to the same constraining action on θ_{xkn} dof.

Moreover, following the same procedure one could evaluate Kirchhoff modes along the other kinematic symmetry line (y -axis)

and arrive at stiffening constraints of type (19a), i.e.,

$$\theta_{yin} + \theta_{yjn} + \theta_{ykn} + \dots \neq 0 \quad (21)$$

Thus, constraints (20a) and (21) are enforced throughout the mesh causing a significant stiffening effect (only few rotational dof remain independent because of these constraints).

Upon examining our numerical solution for mesh A, we find that relations (19a, 20a, 21) are in fact confirmed (e.g., for $L/h=10^4$ we have: $\theta_{xj2}=1.00003\Delta$, $\theta_{xj3}=1.00011\Delta$, $\theta_{xj4}=1.00014\Delta$, $\theta_{xj5}=1.00015\Delta$).

A-type mesh: Clamped edges (1-4) and (3-4). It is a very simple matter to verify, using additional boundary restraints $\theta_{y1}=\theta_{x3}=0$ in (17) and (18), that a single macro-element model will produce a fully locking solution. Furthermore, a multi-element model of this type will also lock (simply apply the additional restraints $\theta_{y1}=\theta_{y3}=\theta_{y4}=\dots=0$ to (20a)). The numerical solutions for mesh A (Table 1) are seen to be in exact agreement with our closed form predictions.

B-type mesh: Simply supported edges (3-4) and (2-3) and boundary restraints $w_2=w_3=w_4=0$, $x_1=x_2=x_3=y_1=y_3=y_4=0$.

Employing these boundary restraints in (15) and (16) gives

$$w_1 + (a/2)\theta_{y2}, \quad \theta_{y2} \neq 0 \quad (22a)$$

$$\kappa_x = \theta_{y2}/a, \quad \kappa_y = 0, \quad \kappa_{xy} = -\theta_{y2}/a \quad (22b)$$

$$\theta_{x4} \neq 0, \quad w_1 + (a/2)\theta_{x4} \quad (23a)$$

$$\kappa_x = 0, \quad \kappa_y = \theta_{x4}/a, \quad \kappa_{xy} = -\theta_{x4}/a \quad (23b)$$

Clearly, the vanishing θ_{y2} and θ_{x4} dof are spurious, and consequently, each element in this mesh will lock (all dof and curvatures vanish, i.e., we have a trivial solution).

We can also verify analytically whether a finer mesh of type B has a chance to avoid spurious locking. To do this, let us consider elements P_n and Q_n ($n=1,2,\dots,N$) along a symmetry axis x (refer to Fig. 6b). The linear Kirchhoff modes for these elements are

$$\theta_{xi(n+1)} - \theta_{xin} - \theta_{yin} + \theta_{yjn} = 0 \quad (P_n) \quad (24)$$

$$\theta_{xjn} - \theta_{xj(n+1)} + \theta_{yi(n+1)} - \theta_{yj(n+1)} = 0 \quad (Q_n) \quad (25)$$

Owing to the kinematic symmetry along x -axis ($\theta_{xin}=0$ for all n)

constraint (24) takes a spurious form

$$\theta_{yin} + \theta_{yjn} \quad (24a)$$

and using (24a) in (25) we obtain another set of spurious constraints on Q_n elements

$$\theta_{xjn} + \theta_{xj(n+1)} \quad (25a)$$

Now, with (i_1-j_1) edge under simply supported restraints (i.e., $\theta_{xj1}=0$), (25a) becomes

$$\theta_{xj(N+1)} + \theta_{xjN} + \theta_{xj(N-1)} + \dots + \theta_{xj2} + \theta_{xj1} = 0 \quad (25b)$$

Obviously, the adjacent row of elements will suffer from the same spurious constraints. Thus, no matter what the boundary condition on the last row of elements, all θ_x dof will vanish.

By considering elements along the other symmetry line (y-axis) we determine that all θ_y dof also vanish.

It is then clear that a B-type mesh will lock under simply supported (and clamped) restraints regardless of how fine the discretization might be. (The fact that both θ_x and θ_y are forced to vanish implies zero bending energy. The deflection, w , will also vanish because of its dependence upon rotational dof via constant Kirchhoff constraints $A_\theta, B_\theta \neq 0$).

REMARK 4.1 It must be kept in mind that for a given plate thickness, however small, as the number of elements approaches a large value, the element penalty parameter will no longer be large. In this case, Kirchhoff modes will not be enforced on the element level, and we should have a nonlocking solution at hand.

4.2 C-type mesh. A typical macro-element, C-type discretization for a symmetric plate quadrant is shown in Fig. 5b. For each of the four MIN3_{*} elements, comprising the model, we can write the element Kirchhoff modes (13b) in an explicit algebraic form which already incorporates the necessary boundary restraints.

Simply supported edges (2-3) and (3-4) and boundary restraints
 $\underline{w_2=w_3=w_4=\theta_{x1}=\theta_{x2}=\theta_{x3}=\theta_{y1}=\theta_{y3}=\theta_{y4}=0.}$

The Kirchhoff modes for the four MIN3_{*} elements are
element R

$$\begin{aligned} w_1 &+ (a/2) \theta_{y2}, \\ 2w_5 - w_1 &+ (a/2) (\theta_{y2} - \theta_{x5} - \theta_{y5}) \\ \theta_{y2} &+ 2\theta_{y5} \end{aligned} \quad (26)$$

element S

$$\begin{aligned} 2w_5 + (a/2)(\theta_{y2} - \theta_{x5} + \theta_{y5}), \\ \theta_{y2} + 2\theta_{x5} \end{aligned} \quad (27)$$

element P

$$\begin{aligned} \theta_{x4} + 2\theta_{y5}, \\ 2w_5 + (a/2)(\theta_{x4} + \theta_{x5} - \theta_{y5}) \end{aligned} \quad (28)$$

element Q

$$\begin{aligned} w_1 - 2w_5 + (a/2)(\theta_{x5} + \theta_{y5} - \theta_{x4}), \\ \theta_{x4} + 2\theta_{x5}, \\ w_1 + (a/2)\theta_{x4} \end{aligned} \quad (29)$$

Solving (26-29) simultaneously yields the following constraints:

$$\begin{aligned} w_1 + (a/2)\theta_{y2}, \quad w_5 + (a/4)\theta_{y2}, \\ \theta_{x5} = \theta_{y5}, \quad \theta_{x4} = \theta_{y2}, \quad \theta_{y2} + 2\theta_{y5} \end{aligned} \quad (30)$$

Equations (30) clearly show a nonlocking nature of the solution, i.e., the dof that are not restrained by boundary conditions remain finite (nonzero). Intuitively, one could also expect that a larger mesh of this type would not engender locking, and this, in fact, is confirmed by our numerical solution for mesh C (refer to Fig. 3 and Table 1).

Clamped edges. To achieve a clamped condition, we simply enforce two additional boundary restraints $\theta_{y2} = \theta_{x4} = 0$ in (30). In this case, all dof vanish, and the mesh is fully locked.

In contrast to the A- and B-type discretizations, where the behavior of a large mesh follows the pattern established by a single macro-element model, the C-type mesh (for a clamped plate) does not exhibit overall locking once we go to 2x2 and higher refinement levels.

To verify this analytically, without performing the finite element numerical computations, we can proceed starting with elements that form a clamped corner of the plate (i.e., elements that are most severely restrained). For example, let us consider a 2x2 symmetric quadrant discretization depicted in Fig. 6c. Here, edges (k_1-k_3) and (i_1-k_1) are clamped, while (i_1-i_3) and (i_3-k_3) are kinematic symmetry axes. Taking into account the boundary restraints imposed on elements S_1 and P_1 (i.e.,

$w_r = \theta_{xr} = \theta_{yr} = 0$, where $r = k_1, k_2, j_1$, the Kirchhoff modes become

element S_1

$$w_{n1} \rightarrow (a/4)(\theta_{yn1} - 3\theta_{xn1}), \quad \theta_{xn1} \rightarrow 0 \quad (31)$$

element P_1

$$w_{n1} \rightarrow (a/4)(\theta_{xn1} - 3\theta_{yn1}), \quad \theta_{yn1} \rightarrow 0 \quad (32)$$

which result in $w_{n1} \rightarrow 0$ or locking for the two elements. This solution plus the boundary restraints at nodes j_1 and k_2 can be used to compute shear constraints for elements R_1 and Q_1 , which happened to be identical for the two elements, i.e.,

elements Q_1 and R_1

$$\begin{aligned} w_{j2} &\rightarrow (a/2)(2\theta_{yj2} - \theta_{xj2}), \\ \theta_{xj2} &\rightarrow \theta_{yj2}, \\ w_{j2} &\rightarrow (a/2)(2\theta_{xj2} - \theta_{yj2}) \end{aligned} \quad (33)$$

Evidently, there is no locking in either of the elements.

We proceed further along the same lines and examine the adjacent elements S_2 and P_2 . Thus, we obtain

element S_2

$$\begin{aligned} w_{m1} &\rightarrow (a/4)(\theta_{ym1} - \theta_{xm1}), \\ \theta_{xm1} &\rightarrow 0 \end{aligned} \quad (34)$$

element P_2

$$\begin{aligned} w_{j2} &\rightarrow (a/2)(2\theta_{ym1} - \theta_{xj2}), \\ \theta_{xj2} &\rightarrow 2\theta_{ym1} - \theta_{yj2}, \\ 2w_{m1} - w_{j2} &\rightarrow (a/2)(2\theta_{xj2} + \theta_{xm1} + \theta_{yj2} - 3\theta_{ym1}) \end{aligned} \quad (35)$$

The second of constraints (34) is spurious. In this case, however, locking is only of a local nature since the other dof remain nonzero.

Similarly, one could examine the rest of the mesh and find the second spurious constraint, namely, $\theta_{yn2} \rightarrow 0$, which is also only local. Thus, the mesh possesses two spurious constraints that do not trigger a global (mesh-wide) locking. The same kind of a local locking trend is achieved with a larger mesh, namely,

the 4x4 C-discretization of Fig. 4. Note that the resulting solution has a substantially greater error than that in the simply supported case (refer to Table 1). This is obviously due to local locking of the clamped plate.

5. ALTERNATIVE BOUNDARY DISCRETIZATIONS. From the previous discussion it becomes apparent that boundary restraints, element geometry, and orientation are the contributing factors that affect the condition of the shear strains. Shear locking and overall solution stiffening were shown to emanate from spurious Kirchhoff modes. In the following two examples with uniformly loaded square plates we demonstrate how simple changes in the boundary element orientations and geometry change the character of a solution.

5.1 Simply supported plates. In Fig. 7 are depicted five meshes for a symmetric plate quadrant which represent slight boundary element modifications of meshes A and B. The results are summarized in Table 2. Interestingly, the global locking experienced by mesh B is no longer present in these alternative discretizations. However, four of these meshes exhibit local locking of some rotational dof which are responsible for the overall solution error.

5.2 Clamped plates. Discretizations G and H of Fig. 8 were analyzed under clamped boundary restraints. Neither mesh exhibits global locking (refer to Table 3), however, local locking takes place at several nodes.

6. CONCLUDING REMARKS. In this paper we elaborated upon an a priori, element-by-element algebraic procedure for identifying such plate modeling shortcomings as shear locking and/or related solution stiffening. Closed form solutions and numerical computations were carried out for several mesh patterns and boundary conditions using a conforming, three-node Mindlin element (MIN3_{*}). The element tends to lock occasionally and, in some cases, produces an overall plate-stiffening effect. These pitfalls were shown to emanate from spurious Kirchhoff modes which arise from overrestraining of the element kinematic field by boundary conditions. The finite element solutions were found to be in complete agreement with the analytic a priori predictions.

The enhanced version of the element (MIN3) uses a finite element shear correction device, which effectively relaxes the enforcement of spurious Kirchhoff modes. The approach also ensures a well-conditioned element stiffness over the entire range of A/h^2 ratios, and there exists no shear locking to hinder MIN3's performance. (Note that MIN3_{*}'s stiffness becomes ill-conditioned for very large values of A/h^2 , thus requiring large computer-word lengths to avoid conditioning errors). For the purpose of comparison, the MIN3 results for the square plate problems of section 3 are summarized in Tables 4 and 5. (Several

important test problems solved with this element were reported in [29]).

Finally, it is worth commenting on the general character of the shear (or penalty) constraints discussed herein. The mathematical form of these constraints is virtually identical (or closely related) to those of the penalty formulations for incompressible elasticity, plasticity and fluid mechanics. The locking pitfall, often hindering such penalty methods, is the consequence of deficient penalty strain approximations, which often are subject to boundary restraints and, therefore, further constraining of the spurious nature. A close examination of penalty modes, in the manner presented here for the Kirchhoff constraints, may be seen as a necessary prerequisite to understanding element behavior. Such an analysis may also point toward new and more effective ways of improving these penalty methods.

REFERENCES.

1. G. Wempner, J. T. Oden and D. Kross, "Finite element analysis of thin shells," Proceedings of ASCE, J. Engrg. Mech. Div. 94 (1968) 1273-1294.
2. J. A. Stricklin, W. E. Haisler, P. R. Tisdale and R. Gunderson, "A rapidly converging triangular plate bending element," AIAA J. 7 (1969) 180-181.
3. J. L. Batoz, K. J. Bathe and L. W. Ho, "A study of three-node triangular plate bending elements," Internat. J. Numer. Meths. Engrg. 15 (1980) 1771-1812.
4. J. L. Batoz, "An explicit formulation for an efficient triangular plate-bending element," Internat. J. Numer. Meths. Engrg. 18 (1982) 1077-1089.
5. A. Needleman and C. F. Shih, "A finite element method for plane strain deformations of incompressible solids," Comput. Meths. Appl. Mech. Engrg. 15 (1978) 223-240.
6. O. C. Zienkiewicz, R. L. Taylor and J. M. Too, "Reduced integration techniques in general analysis of plates and shells," Internat. J. Numer. Meths. Engrg. 3 (1971) 275-290.
7. T. J. R. Hughes, R. L. Taylor and W. Kanoknukulchai, "A simple and efficient element for plate bending, Internat. J. Numer. Meths. Engrg. 11 (1977) 1529-1543.
8. O. C. Zienkiewicz, J. Bauer, K. Morgan and E. Onate, "A simple and efficient element for axisymmetric shells," Internat. J. Numer. Meths. Engrg. 11 (1977) 1545-1558.

9. E. D. L. Pugh, E. Hinton, and O. C. Zienkiewicz, "A study of quadrilateral plate bending elements with 'reduced' integration," *Internat. J. Numer. Meths. Engrg* 12 (1978) 1059-1079.
10. D. S. Malkus and T. J. R. Hughes, "Mixed finite-element methods--Reduced and selective reduced integration techniques: a unification of concepts," *Comput. Meths. Appl. Mech. Engrg.* 15 (1978) 63-81.
11. T. J. R. Hughes and M. Cohen, "The 'Heterosis' finite element for plate bending," *Comput. & Structures* 9 (1978) 445-450.
12. T. J. R. Hughes, M. Cohen and M. Haroun, "Reduced and selective integration techniques in the finite element analysis of plates," *Nucl. Engrg. Design* 46 (1978) 203-222.
13. R. H. MacNeal, "A simple quadrilateral shell element," *Computers and Structures* 9 (1978) 175-183.
14. T. J. R. Hughes, W. K. Liu and A. Brooks, "Review of finite element analysis of incompressible viscous flows by the penalty function formulation," *J. Comp. Phys.* 30 (1979) 1-60.
15. T. J. R. Hughes and T. E. Tezduyar, "Finite elements based upon Mindlin plate theory with particular reference to the four-node bilinear isoparametric element," *ASME J. Appl. Mech.* 48 (1981) 587-596.
16. A. K. Noor and J. M. Peters, "Mixed models and reduced/selective integration displacement models for nonlinear analysis of curved beams," *Internat. J. Numer. Meths. Engrg.* 17 (1981) 615-631.
17. T. J. R. Hughes and R. L. Taylor, "The linear triangular bending element," in *IV-MAFELAP 1981* (Ed. J. R. Whiteman), Academic Press, London (1982) 127-142.
18. R. H. MacNeal, "Derivation of element stiffness matrices by assumed strain distributions," *Nucl. Engrg. Design* 70 (1982) 3-12.
19. T. J. Oden, N. Kikuchi and Y. J. Song, "Penalty finite element methods for the analysis of Stokesian flows," *Comput. Meths. Appl. Mech. Engrg.* (1982) 297-329.
20. M. A. Crisfield, "A four-node thin bending element using shear constraints -- a modified version of Lyons' element," *Comput. Meths. Appl. Mech. Engrg* 38 (1983) 93-120.

21. H. Stolarski and T. Belytschko, "Shear and membrane locking in curved C^0 elements," Comput. Meths. Appl. Mech. Engrg. 41 (1983) 279-296.
22. T. Belytschko, H. Stolarski and N. Carpenter, "A C^0 triangular plate element with one-point quadrature," Internat. J. Numer. Meths. Engrg. 20 (1984) 787-802.
23. A. Tessler and S. B. Dong, "On a hierarchy of conforming Timoshenko beam elements," Computers and Structures, 14 (1981) 335-344.
24. A. Tessler, "An efficient, conforming axisymmetric shell element including transverse shear and rotary inertia," Computers and Structures 15 (1982) 567-574.
25. A. Tessler, "On a conforming, Mindlin-type plate element," in IV-MAFELAP 1981 (ed., J. R. Whiteman), Academic Press, London (1982) 119-126.
26. I. Fried, "Shear in C^0 and C^1 plate bending elements," Internat. J. Solids and Structures 9 (1973) 449-460.
27. I. Fried, "Residual energy balancing technique in the generation of plate bending finite element, Computers & Structures 4 (1974) 771-778.
28. A. Tessler and T. J. R. Hughes, "An improved treatment of transverse shear in the Mindlin-type four-node quadrilateral element," Comput. Meths. Appl. Mech. Engrg. 39 (1983) 311-335.
29. A. Tessler and T. J. R. Hughes, "A three-node Mindlin plate element with improved transverse shear," (to appear in Computer Meths. Appl. Mech. Engrg., 1985).
30. I. Fried, A. Johnson and A. Tessler, "Minimal degree thin triangular plate bending finite elements of order two and four" (to appear in Compt. Methds. Appl. Mech. Engrg., 1985)
31. K. C. Park and D. L. Flagg, "An operational procedure for the symbolic analysis of the finite element method," Comput. Meth. Appl. Mech. Engrg. 42 (1984) 37-46.
32. K. C. Park and D. L. Flagg, "A Fourier analysis of spurious mechanisms and locking in the finite element method," Comput. Meths. Appl. Mech. Engrg. 46 (1984) 65-81.
33. R. D. Mindlin, "Influence of rotatory inertia and shear on flexural motions of isotropic, elastic plates," ASME J. Appl. Mech. 18 (1951) 31-38.
34. S. Timoshenko and S. Woinowsky-Krieger, Theory of Plates and Shells, McGraw-Hill Book Co., New York (1959)

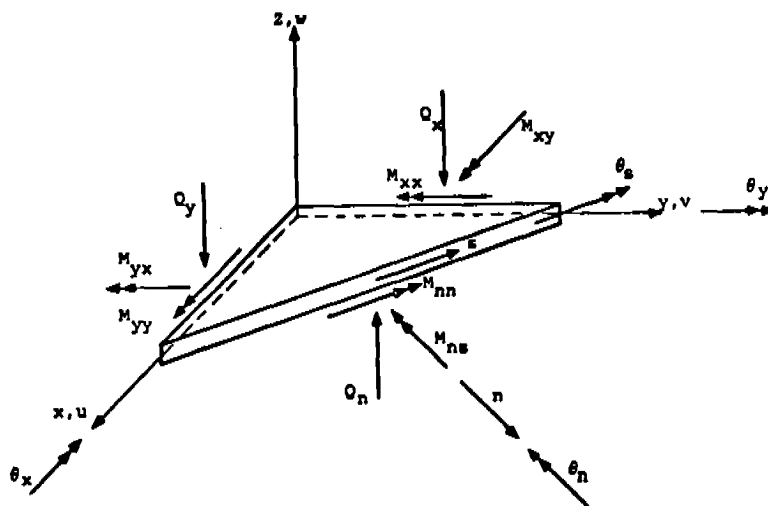


Fig. 1. Plate notations for kinematic variables and stress resultants.

SHAPE FUNCTIONS		INITIAL NODAL CONFIGURATION	CONTINUOUS SHEAR EDGE CONSTRAINTS: $(w_s + \theta_n)_{,s} = 0$	CONSTRAINED NODAL CONFIGURATION
w	θ_x, θ_y			
QUADRATIC	LINEAR		THREE EDGE CONSTRAINTS 	'MIN3'

KEY:

- w, θ_x, θ_y DEGREES OF FREEDOM
- w DEGREES OF FREEDOM

Fig. 2. Triangular element configuration.

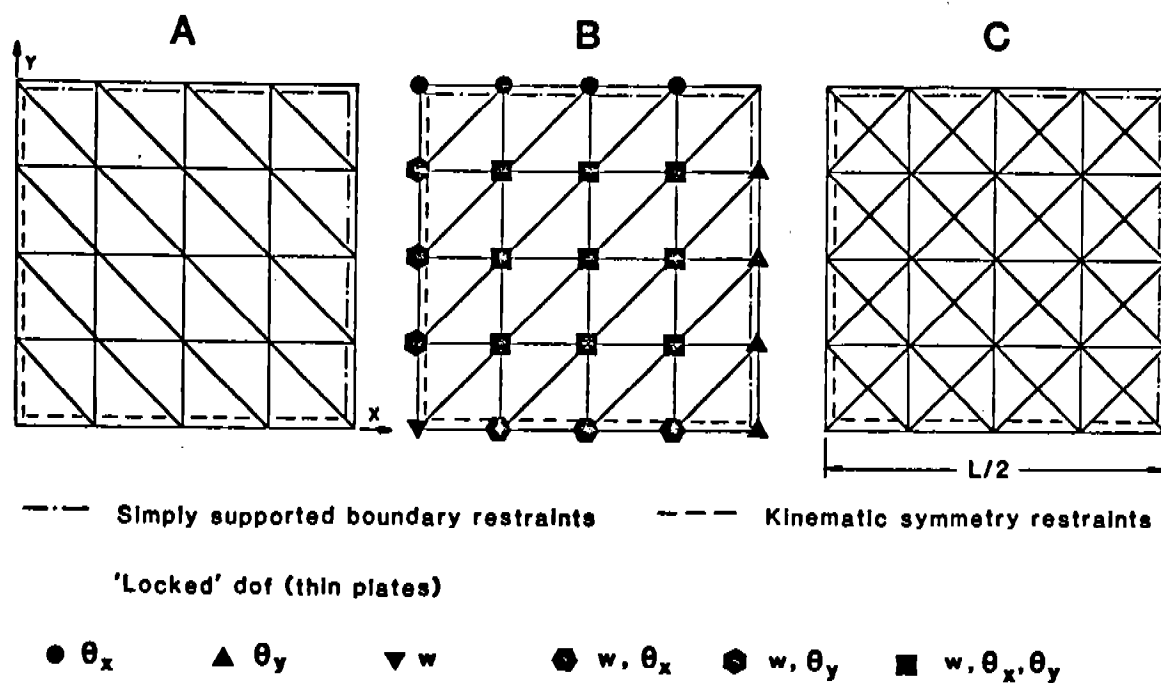


Fig. 3. Simply supported square plates: standard symmetric-quadrant meshes.

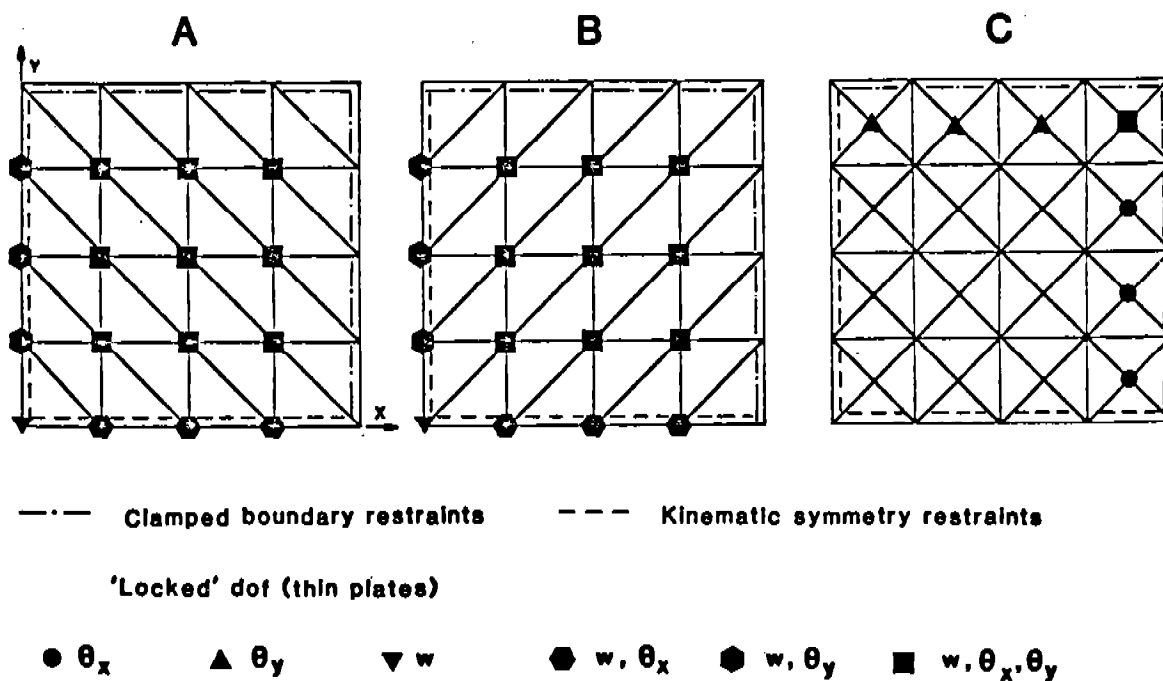
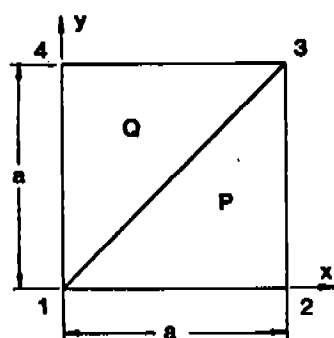


Fig. 4. Clamped square plates: standard symmetric-quadrant meshes.

a. A & B-type macro-element



b. C-type macro-element

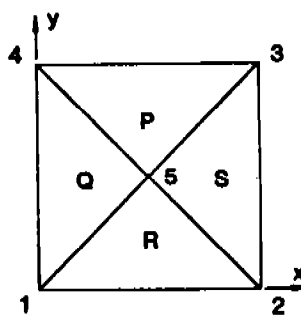
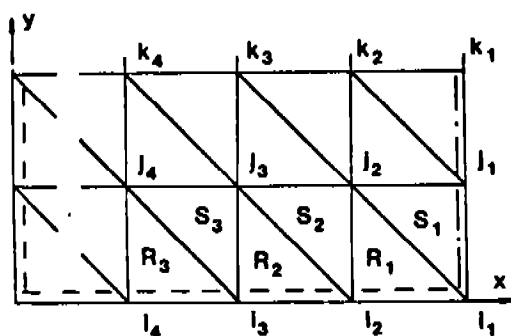
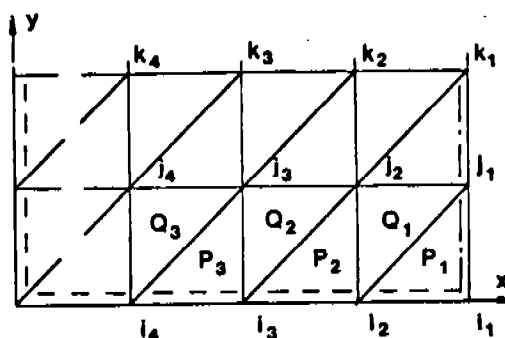


Fig. 5. Macro-element meshes.

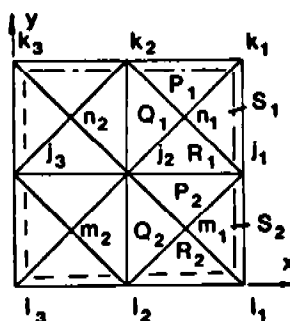
a. A-type mesh fragment



b. B-type mesh fragment



c. 2x2 C-type mesh



--- Boundary restraints
 --- Kinematic symmetry restraints

Fig. 6. Fragments of standard meshes.

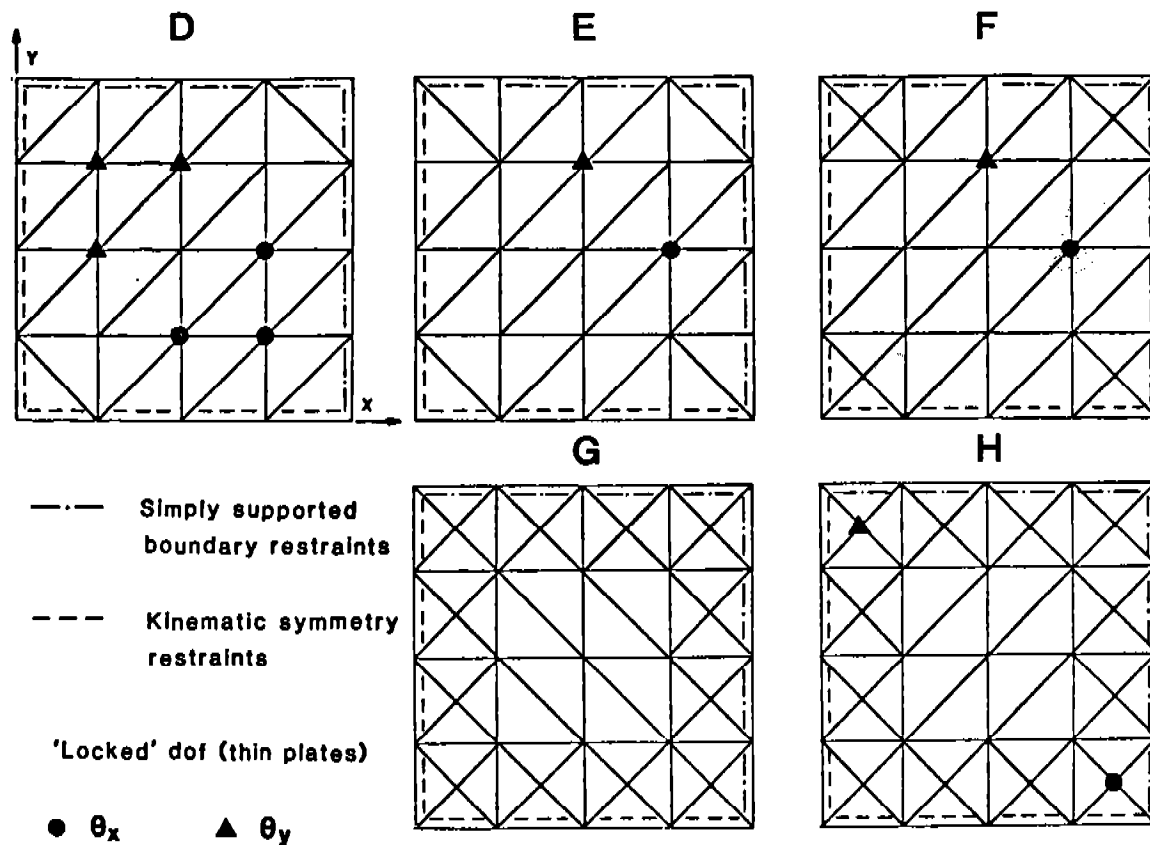


Fig. 7. Simply supported square plates: alternative symmetric-quadrant meshes.

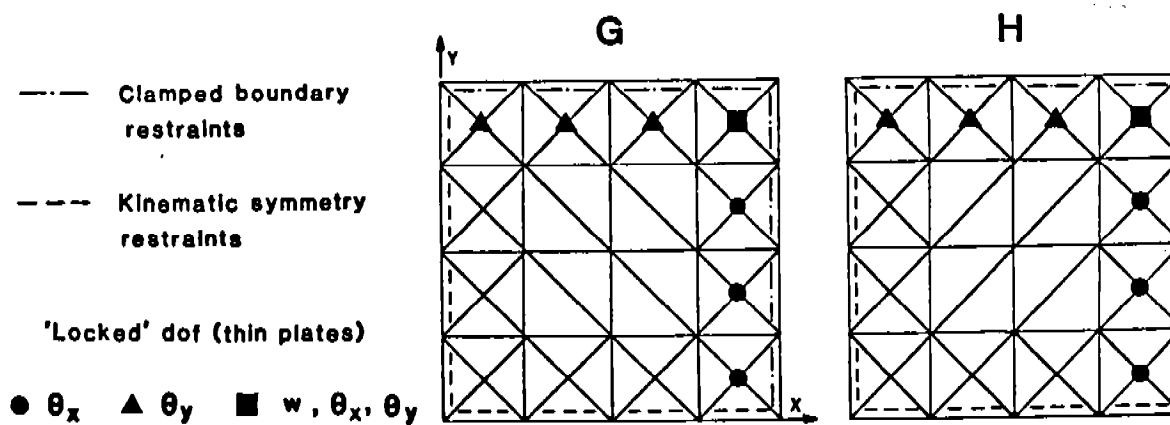


Fig. 8. Clamped square plates: alternative symmetric-quadrant meshes.

Table 1. Simply supported and clamped square plates under uniform load; MIN3, center deflection results normalized with respect to analytic solution according to Mindlin theory ($E=10.92 \times 10^6$, $\nu=0.3$).

L/h	simply supported			clamped		
	mesh					
	A	B	C	A	B	C
10	0.981	0.970	0.993	0.909	0.947	0.973
10 ²	0.891	0.728	0.990	0.504	0.557	0.935
10 ⁴	0.801	3.8x10 ⁻⁴	0.989	1.2x10 ⁻⁴	1.5x10 ⁻⁴	0.931
10 ⁶	0.801	3.8x10 ⁻⁶	0.989	1.2x10 ⁻⁶	1.5x10 ⁻⁶	0.931

Table 2. Simply supported square plates under uniform load; MIN3, center deflection results normalized with respect to analytic solution according to Mindlin theory ($E=10.92 \times 10^6$, $\nu=0.3$).

L/h	mesh				
	D	E	F	G	H
10	0.971	0.971	0.975	0.992	0.988
10 ²	0.801	0.830	0.831	0.986	0.955
10 ⁴	0.540	0.579	0.577	0.984	0.931
10 ⁶	0.540	0.579	0.577	0.984	0.931

Table 3. Clamped square plates under uniform load;
MIN3 center deflection results normalized with
respect to analytic solution according to Mindlin
theory ($E=10.92 \times 10^6$, $\nu=0.3$).

L/h	mesh	
	G	H
10	0.966	0.968
10 ²	0.922	0.942
10 ⁴	0.915	0.968
10 ⁶	0.915	0.968

Table 4. Simply supported square plates under uniform load;
MIN3 center deflection results normalized with
respect to analytic solution according to Mindlin
theory ($E=10.92 \times 10^6$, $\nu=0.3$).

L/h	mesh				
	D	E	F	G	H
10	0.997	0.997	0.999	1.007	1.003
10 ²	0.992	0.991	0.992	1.006	1.001
10 ⁴	0.991	0.991	0.991	1.006	1.000
10 ⁶	0.991	0.991	0.991	1.006	1.000

Table 5. Clamped square plates under uniform load;
MIN3 center deflection results normalized with
respect to analytic solution according to Mindlin
theory ($E=10.92 \times 10^6$, $\nu=0.3$).

L/h	mesh	
	G	H
10	1.011	1.014
10 ²	1.003	1.003
10 ⁴	1.004	1.001
10 ⁶	1.004	1.001

ON THE SOLVABILITY AND COMPUTATIONAL ASPECTS OF A REFINED SHEAR DEFORMATION PLATE THEORY

J. N. Reddy
Department of Engineering Science and Mechanics
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061 USA

ABSTRACT

The existence and uniqueness of solutions to the third-order shear deformation theory developed recently by the author are examined, and the computational aspects of the associated finite-element models are discussed. The existence and uniqueness of solutions of the linear theory are proved using the Lax-Milgram theorem. A new mixed finite-element model that uses displacements and bending moments as degrees of freedom and a displacement finite-element model are discussed. The mixed element is a C^0 element while the displacement element is a C^1 element.

1. INTRODUCTION

The origin of displacement-based theories is due to Basset [1]. In a displacement-based theory, it is assumed that the displacement components of the plate can be expanded in series of powers of the thickness coordinate z . For example, the displacement component u in the x -direction in the N -th order theory is written in the form

$$u(x,y,z) = u^0(x,y) + \sum_{n=1}^N z^n u^{(n)}(x,y) \quad (1)$$

where x and y are the Cartesian coordinates in the middle plane of the plate, and $u^{(n)}$ have the meaning

$$u^{(n)}(x,y) = \frac{d^n u}{dz^n} \Big|_{z=0}, \quad n = 0, 1, 2, \dots \quad (2)$$

Basset's work did not receive as much attention as it deserves. In 1949

NACA technical note, Hildebrand, Reissner and Thomas [2] presented (also see Hencky [3]) a first-order shear deformation theory [a special case of Eq. (1)],

$$\begin{aligned}u(x,y,z) &= u^0(x,y) + z\psi_x(x,y) \\v(x,y,z) &= v^0(x,y) + z\psi_y(x,y) \\w(x,y,z) &= w^0(x,y)\end{aligned}\tag{3}$$

The differential equations of the theory were derived using the principle of the minimum total potential energy. This leads to five equilibrium equations in the five generalized displacement variables, u^0, v^0, w^0, ψ_x and ψ_y . Mindlin [4] presented a dynamic analysis of Hencky's theory [3], and used the displacement field in Eq. (3) for the vibration of isotropic plates. We shall refer to the shear deformation theory based on the displacement field (3) as the first-order shear deformation theory, although it is known as the Mindlin or Reissner-Mindlin theory in the literature. For additional references on the history of the shear deformation theory, the reader is advised to consult Reference 5.

Recently, Levinson [6] and Murthy [7] presented third-order theories in which vanishing of the transverse shear stresses on the bounding planes of the plate and transverse inextensibility of the normals were assumed. However, both authors used the equilibrium equations of the first-order theory in their analysis, and thus the higher-order terms of the displacement field were accounted only in the calculation of the strains but not in the governing differential equations or in the boundary conditions. Recently, the present author (see [8,9]) derived variationally consistent equations of motion by means of Hamilton's principle. The theory presented in [8] accounts for

moderately large deflections but is limited to orthotropic plates, while that in [9] is for the small-deflection theory of laminated plates. Finite element models of these theories were presented by Reddy and his colleagues [10,11]. The present paper contains a review of the theory, some mathematical results on the existence and uniqueness of solutions of the linear theory, and a discussion of the displacement and mixed finite element models.

2. EQUATIONS OF THE HIGHER-ORDER THEORY

Consider a rectangular laminate with planeform dimensions a and b and thickness h . The coordinate system is taken such that the x - y plane coincides with the mid-plane of the plate, and the z -axis is perpendicular to that plane. The plate is composed of perfectly bonded orthotropic layers with the principal material axes of each layer oriented arbitrarily with respect to the plate axes.

In order to obtain a parabolic distribution of the transverse shear stresses, a cubic expansion is used for the inplane displacements and the transverse inextensibility (i.e., the transverse normal strain ϵ_z is zero) is assumed. The resulting displacement is given by (see Reddy [8,9]):

$$\begin{aligned} u_1(x,y,z) &= u(x,y) + z\left[\psi_x(x,y) - \frac{4}{3}\left(\frac{z}{h}\right)^2\left(\psi_x(x,y) + \frac{\partial w}{\partial x}(x,y)\right)\right] \\ u_2(x,y,z) &= v(x,y) + z\left[\psi_y(x,y) - \frac{4}{3}\left(\frac{z}{h}\right)^2\left(\psi_y(x,y) + \frac{\partial w}{\partial y}(x,y)\right)\right] \\ u_3(x,y,z) &= w(x,y) \quad , \end{aligned} \tag{4}$$

where u_1 , u_2 and u_3 are the displacements in the x -, y - and z -directions, respectively. The displacements of a point (x,y) on the midplane of the plate are denoted by u , v and w ; ψ_x and ψ_y are the

rotations of normals to the midplane about the y- and x-axes, respectively. The strains can be obtained from Eq. (4).

Assuming that each layer of the laminate possesses a plane of material symmetry parallel to the x-y plane, the constitutive equations for the k-th layer can be written as

$$\begin{Bmatrix} \bar{\sigma}_1 \\ \bar{\sigma}_2 \\ \bar{\sigma}_6 \end{Bmatrix}^{(k)} = \begin{bmatrix} \bar{Q}_{11} & \bar{Q}_{12} & 0 \\ \bar{Q}_{12} & \bar{Q}_{22} & 0 \\ 0 & 0 & \bar{Q}_{66} \end{bmatrix}^{(k)} \begin{Bmatrix} \bar{\epsilon}_1 \\ \bar{\epsilon}_2 \\ \bar{\epsilon}_6 \end{Bmatrix}^{(k)},$$

$$\begin{Bmatrix} \bar{\sigma}_4 \\ \bar{\sigma}_5 \end{Bmatrix}^{(k)} = \begin{bmatrix} \bar{Q}_{44} & 0 \\ 0 & \bar{Q}_{55} \end{bmatrix}^{(k)} \begin{Bmatrix} \bar{\epsilon}_4 \\ \bar{\epsilon}_5 \end{Bmatrix}^{(k)} \quad (5)$$

where $\bar{\sigma}_i$ and $\bar{\epsilon}_i$ ($i = 1, 2, 4, 5, 6$) are the stress and strain components referred to lamina coordinates and \bar{Q}_{ij} 's are the plane-stress reduced elastic constants in the material axes of the k-th lamina,

$$\bar{Q}_{11} = E_1 / (1 - \nu_{12}\nu_{21}), \quad \bar{Q}_{22} = \frac{E_2}{E_1} \bar{Q}_{11}, \quad \bar{Q}_{12} = \nu_{12} \bar{Q}_{22}$$

$$\bar{Q}_{44} = G_{23}, \quad \bar{Q}_{55} = G_{13}, \quad \bar{Q}_{66} = G_{12} \quad (6)$$

and E_i , ν_{ij} and G_{ij} are the usual engineering constants.

The equations of equilibrium can be obtained using the principle of virtual displacements. In analytical form, the principle can be stated as follows (see [12]):

$$\int_V (\delta U + \delta W) dV = 0 \quad (7)$$

where U is the total strain energy due to deformation, W is the potential of external loads, and V is the volume of the laminate and δ denotes the variational symbol. The following five equations are obtained for the linear case:

$$N_{1,x} + N_{6,y} = 0$$

$$N_{6,x} + N_{2,y} = 0$$

$$Q_{1,x} + Q_{2,y} - \frac{4}{h^2} (R_{1,x} + R_{2,y}) + \frac{4}{3h^2} (P_{1,xx} + 2P_{6,xy} + P_{2,yy}) = q$$

$$M_{1,x} + M_{6,y} - Q_1 + \frac{4}{h^2} R_1 - \frac{4}{3h^2} (P_{1,x} + P_{6,y}) = 0$$

$$M_{6,x} + M_{2,y} - Q_2 + \frac{4}{h^2} R_2 - \frac{4}{3h^2} (P_{6,x} + P_{2,y}) = 0 \quad (8)$$

where

$$(N_i, M_i, P_i) = \int_{-h/2}^{h/2} \sigma_i(1, z, z^3) dz \quad (i=1, 2, 6)$$

$$(Q_2, R_2) = \int_{-h/2}^{h/2} \sigma_4(1, z^2) dz \quad (9)$$

$$(Q_1, R_1) = \int_{-h/2}^{h/2} \sigma_5(1, z^2) dz$$

In the general case of arbitrary geometry, boundary conditions and lamination scheme, the exact analytical solutions to the set of differential equations in Eq. (8) cannot be found. However, closed-form solutions exist for certain 'simply supported' rectangular plates with two sets of laminate stiffnesses, as described in [8,9,12].

3. EXISTENCE AND UNIQUENESS OF SOLUTIONS

A formal proof of the existence and uniqueness of solutions of the equations of the higher-order theory requires us to establish the positive-definiteness of the bilinear form in an appropriate function

(or vector) space. First, some mathematical preliminars are in order (see [13]).

Let Ω denote the open bounded region of the plate (i.e., the midplane of the plate) and Γ be its boundary. We use the standard notation:

$L_2(\Omega)$ = the complete named (i.e., Hilbert) space of generalized functions that are square integrable in Ω ; i.e., $u \in L_2(\Omega)$ means

$$\int_{\Omega} |u(x,y)|^2 dx dy < \infty$$

$H^1(\Omega)$ = the complete normed space of functions which along with their generalized first derivatives belong to $L_2(\Omega)$; i.e., $u \in H^1(\Omega)$ implies

$$\int_{\Omega} |u(x,y)|^2 dx dy < \infty, \quad \int_{\Omega} \left| \frac{\partial u}{\partial x}(x,y) \right|^2 dx dy < \infty, \quad \int_{\Omega} \left| \frac{\partial u}{\partial y}(x,y) \right|^2 dx dy < \infty$$

$H^2(\Omega)$ = the complete normed space of functions which along with their generalized derivatives up to and including order 2 are in $L_2(\Omega)$

(10a)

We shall use $H_0^1(\Omega)$ to denote the space of functions u from $H^1(\Omega)$ that vanish on the boundary Γ of Ω , and $H_0^2(\Omega)$ to denote the space of functions u from $H^2(\Omega)$ which, along with their first derivatives, vanish on Γ . The norm in $H^m(\Omega)$ [$H^0(\Omega) \equiv L_2(\Omega)$] is given by

$$\|u\|_m^2 = \int_{\Omega} \sum_{i,j=0}^m \left| \frac{\partial^{i+j} u}{\partial x^i \partial y^j} \right|^2 dx dy \quad i+j \leq m \quad (10b)$$

Equations (7) and (8) can be expressed in terms of the five generalized displacements u , v , w , ψ_x and ψ_y . In terms of the displacements, Eq. (7) has the form (for symmetric laminates only):

$$B(\bar{\Lambda}, \Lambda) = \mathfrak{L}(\bar{\Lambda}) \quad (11)$$

where $\bar{\Lambda} = (\bar{w}, \bar{\psi}_x, \bar{\psi}_y)$, $\Lambda = (w, \psi_x, \psi_y)$, and $B(\cdot, \cdot)$ and $\ell(\cdot)$ are the bilinear and linear forms on the vector space, H

$$H = H_0^2(\Omega) \times H_0^1(\Omega) \times H_0^1(\Omega) \quad (12a)$$

The norm in H is defined by

$$\|\Lambda\|_H^2 = \|w\|_2^2 + \|\psi_x\|_1^2 + \|\psi_y\|_1^2 \quad (12b)$$

where $\|\cdot\|_m$ denotes the $H^m(\Omega)$ norm.

In writing Eqs. (11) and (12), it is assumed that the inplane displacements, being uncoupled from w, ψ_x, ψ_y , can be solved independently and that the plate is subjected to clamped boundary conditions. The bilinear form $B(\cdot, \cdot)$ and linear form $\ell(\cdot)$ are given by

$$\begin{aligned} B(\bar{\Lambda}, \Lambda) = & \int_R \left\{ \frac{\partial \bar{\psi}_x}{\partial x} \left[D_{11} \frac{\partial \psi_x}{\partial x} + D_{12} \frac{\partial \psi_y}{\partial y} - \frac{4}{3h^2} F_{11} \left(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2} \right) - \frac{4}{3h^2} F_{12} \left(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2} \right) \right] \right. \\ & + \frac{\partial \psi_y}{\partial y} \left[D_{12} \frac{\partial \psi_x}{\partial x} + D_{22} \frac{\partial \psi_y}{\partial y} - \frac{4}{3h^2} F_{12} \left(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2} \right) - \frac{4}{3h^2} F_{22} \left(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2} \right) \right] \\ & + \left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} \right) \left[D_{66} \left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} \right) - \frac{4}{3h^2} F_{66} \left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} + 2 \frac{\partial^2 w}{\partial x \partial y} \right) \right] \\ & + \left(\psi_x + \frac{\partial \bar{w}}{\partial x} \right) \left(A_{55} - \frac{4}{h^2} D_{55} \right) \left(\psi_x + \frac{\partial w}{\partial x} \right) + \left(\bar{\psi}_y + \frac{\partial \bar{w}}{\partial y} \right) \left(A_{44} - \frac{4}{h^2} D_{44} \right) \left(\psi_y + \frac{\partial w}{\partial y} \right) \\ & - \frac{4}{3h^2} \left(\frac{\partial \bar{\psi}_x}{\partial x} + \frac{\partial^2 \bar{w}}{\partial x^2} \right) \left[F_{11} \frac{\partial \psi_x}{\partial x} + F_{12} \frac{\partial \psi_y}{\partial y} - \frac{4}{3h} H_{11} \left(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2} \right) \right. \\ & - \frac{4}{3h^2} H_{12} \left(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2} \right) \left. - \frac{4}{3h^2} \left(\frac{\partial \bar{\psi}_y}{\partial y} + \frac{\partial^2 \bar{w}}{\partial y^2} \right) \left[F_{12} \frac{\partial \psi_x}{\partial x} + F_{22} \frac{\partial \psi_y}{\partial y} \right. \right. \\ & - \frac{4}{3h^2} H_{12} \left(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2} \right) - \frac{4}{3h^2} H_{22} \left(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2} \right) \left. - \frac{4}{3h^2} \left(\frac{\partial \bar{\psi}_x}{\partial y} + \frac{\partial \bar{\psi}_y}{\partial x} \right) \right. \\ & \left. \left. + 2 \frac{\partial^2 \bar{w}}{\partial x \partial y} \right] \left[F_{66} \left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} \right) - \frac{4}{3h} H_{66} \left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} + 2 \frac{\partial^2 w}{\partial x \partial y} \right) \right] \right\} \end{aligned}$$

$$\begin{aligned}
& - \frac{4}{h^2} (\bar{\psi}_x + \frac{\partial \bar{w}}{\partial x}) (D_{55} - \frac{4}{h^2} F_{55}) (\psi_x + \frac{\partial w}{\partial x}) \\
& - \frac{4}{h^2} (\bar{\psi}_y + \frac{\partial \bar{w}}{\partial y}) (D_{44} - \frac{4}{h^2} F_{44}) (\psi_y + \frac{\partial w}{\partial y}) \} dx dy \quad (13a)
\end{aligned}$$

$$\ell(\bar{\Lambda}) = \int_{\Omega} q \bar{w} dx dy$$

where D_{ij} , F_{ij} and H_{ij} are the laminate stiffnesses,

$$(D_{ij}, F_{ij}, H_{ij}) = \int_{-\frac{h}{2}}^{\frac{h}{2}} Q_{ij}(z^2, z^4, z^6) dz \quad (13b)$$

Equation (11) is called the variational problem associated with Eq. (8) when applied to a clamped laminate. We wish to show that the variational problem (11) has a unique solution in the space H . Toward establishing this result, we use the well-known Lax-Milgram theorem (see Reddy [13]):

Theorem 1: Let H be a Hilbert space, and let $B(\cdot, \cdot)$ be a bilinear form on H with the following properties:

$$(i) \quad |B(\bar{\Lambda}, \Lambda)| \leq M \|\Lambda\|_H \|\bar{\Lambda}\|_H \quad (\text{continuity}) \quad (14a)$$

$$(ii) \quad B(\Lambda, \Lambda) \geq \alpha \|\Lambda\|_H^2 \quad (\text{positive-definiteness}) \quad (14b)$$

for all $\Lambda, \bar{\Lambda} \in H$. Then for any continuous linear functional $\ell(\cdot)$ on H , there exists a unique vector Λ_0 in H such that

$$B(\bar{\Lambda}, \Lambda_0) = \ell(\bar{\Lambda})$$

holds for every $\bar{\Lambda} \in H$.

A proof of the theorem can be found in [13]. Thus, if we can show that $B(\cdot, \cdot)$ and $\ell(\cdot)$ of Eq. (13) are continuous and $B(\cdot, \cdot)$ is positive-definite on H , then we have a unique solution to Eq. (11). It can be shown that the solution of the variational problem is also the solution

of the classical problem (i.e. the original differential equations), provided that the boundary of the domain and the data q are sufficiently smooth.

We now set out to establish the continuity and positive-definiteness properties of $B(\cdot, \cdot)$. The continuity of $\mathfrak{L}(\cdot)$ is obvious.

From Eq. (13a), we have

$$\begin{aligned}
 B(\bar{\Lambda}, \Lambda) = & \int_{\Omega} [c_1 (\frac{\partial \bar{w}}{\partial x} + \bar{\psi}_x) (\frac{\partial w}{\partial x} + \psi_x) + c_2 (\frac{\partial \bar{w}}{\partial y} + \bar{\psi}_y) (\frac{\partial w}{\partial y} + \psi_y) \\
 & + c_3 \frac{\partial^2 \bar{w}}{\partial x^2} \frac{\partial^2 w}{\partial x^2} + c_4 \frac{\partial^2 \bar{w}}{\partial y^2} \frac{\partial^2 w}{\partial y^2} + c_5 (\frac{\partial^2 \bar{w}}{\partial x^2} \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 \bar{w}}{\partial x^2} \frac{\partial^2 w}{\partial y^2}) \\
 & + 4c_6 \frac{\partial^2 \bar{w}}{\partial x \partial y} \frac{\partial^2 w}{\partial x \partial y} + c_7 (\frac{\partial^2 \bar{w}}{\partial x^2} \frac{\partial \psi_x}{\partial x} + \frac{\partial^2 \bar{w}}{\partial x^2} \frac{\partial \bar{\psi}_x}{\partial x}) \\
 & + c_8 (\frac{\partial^2 \bar{w}}{\partial y^2} \frac{\partial \psi_y}{\partial y} + \frac{\partial^2 \bar{w}}{\partial y^2} \frac{\partial \bar{\psi}_y}{\partial y}) + c_9 (\frac{\partial^2 \bar{w}}{\partial x^2} \frac{\partial \psi_y}{\partial y} + \frac{\partial^2 \bar{w}}{\partial y^2} \frac{\partial \psi_x}{\partial x}) \\
 & + \frac{\partial^2 \bar{w}}{\partial x^2} \frac{\partial \bar{\psi}_y}{\partial y} + \frac{\partial^2 \bar{w}}{\partial y^2} \frac{\partial \bar{\psi}_x}{\partial x}) + 2c_{10} [(\frac{\partial \bar{\psi}_x}{\partial y} + \frac{\partial \bar{\psi}_y}{\partial x}) (\frac{\partial^2 w}{\partial x \partial y}) \\
 & + (\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x}) \frac{\partial^2 w}{\partial x \partial y}] + c_{11} \frac{\partial \bar{\psi}_x}{\partial x} \frac{\partial \psi_x}{\partial x} + c_{12} (\frac{\partial \bar{\psi}_x}{\partial x} \frac{\partial \psi_y}{\partial y} + \frac{\partial \psi_x}{\partial x} \frac{\partial \bar{\psi}_y}{\partial y}) \\
 & + c_{13} \frac{\partial \bar{\psi}_y}{\partial y} \frac{\partial \psi_y}{\partial y} + c_{14} (\frac{\partial \bar{\psi}_x}{\partial y} + \frac{\partial \bar{\psi}_y}{\partial x}) (\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x})] dx dy \quad (15)
 \end{aligned}$$

where

$$c_1 = A_{55} + \frac{16}{h^4} F_{55}, \quad c_2 = A_{44} + \frac{16}{h^4} F_{44}$$

$$c_3 = \frac{16}{9h^4} H_{11}, \quad c_4 = \frac{16}{9h^4} H_{22}, \quad c_5 = \frac{16}{9h^4} H_{12}, \quad c_6 = \frac{16}{9h^4} H_{66}$$

$$\begin{aligned}
c_7 &= \frac{16}{9h^4} H_{11} - \frac{4}{3h^2} F_{11}, \quad c_8 = \frac{16}{9h^4} H_{22} - \frac{4}{3h^2} F_{22}, \quad c_9 = \frac{16}{9h^4} H_{12} - \frac{4}{3h^2} F_{12} \\
c_{10} &= \frac{16}{9h^4} H_{66} - \frac{4}{3h^2} F_{66}, \quad c_{11} = D_{11} - \frac{8}{3h^2} F_{11} + \frac{16}{9h^4} H_{11} \\
c_{12} &= D_{12} - \frac{8}{3h^2} F_{12} + \frac{16}{9h^4} H_{12}, \quad c_{13} = D_{22} - \frac{8}{3h^2} F_{22} + \frac{16}{9h^4} H_{22} \\
c_{14} &= D_{66} - \frac{8}{3h^2} F_{66} + \frac{16}{9h^4} H_{66}
\end{aligned} \tag{16}$$

The continuity of $B(\cdot, \cdot)$ can be proved by using Hölder's inequality for sums of integrals [13],

$$\begin{aligned}
\sum_{i=1}^n \int_{\Omega} u_i v_i dx dy &\leq \sum_{i=1}^n \left[\int_{\Omega} |u_i|^2 dx dy \right]^{1/2} \left[\int_{\Omega} |v_i|^2 dx dy \right]^{1/2} \\
&\leq \left(\sum_{i=1}^n \int_{\Omega} |u_i|^2 dx dy \right)^{1/2} \left(\sum_{i=1}^n \int_{\Omega} |v_i|^2 dx dy \right)^{1/2}
\end{aligned}$$

We have

$$\begin{aligned}
|B(\bar{\Lambda}, \Lambda)| &\leq M(\|\bar{w}\|_2^2 + \|\bar{\psi}_x\|_1^2 + \|\bar{\psi}_y\|_2^2) \cdot (\|w\|_2^2 + \|\psi_x\|_2^2 + \|\psi_y\|_2^2)^{1/2} \\
&= M\|\bar{\Lambda}\|_H \|\Lambda\|_H
\end{aligned} \tag{17}$$

where $M = \max(|c_1|, |c_2|, \dots, |c_{14}|)$.

Next consider the bilinear form $B(\Lambda, \Lambda)$:

$$\begin{aligned}
B(\Lambda, \Lambda) &= \int_{\Omega} \{c_1 \left(\frac{\partial w}{\partial x} + \psi_x\right)^2 + c_2 \left(\frac{\partial w}{\partial y} + \psi_y\right)^2 + c_3 \left(\frac{\partial^2 w}{\partial x^2}\right)^2 \\
&\quad + c_4 \left(\frac{\partial^2 w}{\partial y^2}\right)^2 + 2c_5 \frac{\partial^2 w}{\partial x^2} \frac{\partial^2 w}{\partial y^2} + 4c_6 \left(\frac{\partial^2 w}{\partial x \partial y}\right)^2 \\
&\quad + 2c_7 \frac{\partial^2 w}{\partial x^2} \frac{\partial \psi_x}{\partial x} + 2c_8 \frac{\partial^2 w}{\partial y^2} \frac{\partial \psi_y}{\partial y} + 2c_9 \left(\frac{\partial^2 w}{\partial x^2}\right) \frac{\partial \psi_y}{\partial y}
\end{aligned}$$

$$\begin{aligned}
& + \frac{\partial^2 w}{\partial y^2} \frac{\partial \psi_x}{\partial x} + 4c_{10} \left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} \right) \frac{\partial^2 w}{\partial x \partial y} \\
& + c_{11} \left(\frac{\partial \psi_x}{\partial x} \right)^2 + 2c_{12} \frac{\partial \psi_x}{\partial x} \frac{\partial \psi_y}{\partial y} + c_{13} \left(\frac{\partial \psi_y}{\partial y} \right)^2 \\
& + c_{14} \left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} \right)^2 \big] dx dy
\end{aligned} \tag{18}$$

Next we use Korn's inequality (see [13]): For $\Lambda = (u, v) \in H_0^1(\Omega) \times H_0^1(\Omega)$ there exists a constant $c(\Omega)$ (depends only on Ω) such that

$$(\|u\|_1^2 + \|v\|_1^2)^{1/2} \leq c(\Omega) \left\{ \int_{\Omega} \left[2 \left(\frac{\partial u}{\partial x} \right)^2 + 2 \left(\frac{\partial v}{\partial y} \right)^2 + \frac{1}{2} \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right)^2 \right] dx dy \right\}^{1/2} \tag{19}$$

Then we have

$$B(\Lambda, \Lambda) \geq \alpha \|\Lambda\|_H^2$$

where $\alpha > 0$ is a constant that depends on Ω and the constants c_1, c_2, \dots, c_{14} , but is independent of Λ .

Since $B(\cdot, \cdot)$ is continuous and positive-definite on H , the variational problem (11) has a unique solution $\Lambda = (w, \psi_x, \psi_y)$ in H . Although we have given the results here for clamped plate, the results can be extended to simply supported plates and plates with mixed boundary conditions.

4. FINITE ELEMENT MODELS

4.1 Displacement Finite Element Model

It is informative to note that the equations of the present higher-order theory has terms from both the classical and first-order shear deformation theory. Although the number of generalized displacements are the same in the higher-order and first-order theories, the higher-

order theory contains terms involving the second-order derivatives of the transverse deflection in the potential energy expression. Hence, like in the classical plate theory, Hermite interpolation of the transverse deflection is required, while Lagrange interpolation of ψ_x and ψ_y is admissible for the displacement finite element model.

The finite-element model is of the form

$$[K^e]\{\Delta^e\} = \{F^e\} \quad (20)$$

where $[K^e]$ is the element stiffness matrix, $\{\Delta^e\}$ is the vector of the nodal values of the generalized displacements, and $\{F^e\}$ is the force vector that contains contributions from both applied loads and contributions from the boundary of the element (i.e., internal forces). In the present study, the isoparametric rectangular element is used for ψ_x and ψ_y and the four-node Hermite cubic element is used for w .

The element stiffness matrix $[K^e]$ is of the form

$$[K^e] = \begin{bmatrix} [K^{11}] & [K^{12}] & [K^{13}] \\ [K^{12}]^T & [K^{22}] & [K^{23}] \\ [K^{13}]^T & [K^{23}]^T & [K^{33}] \end{bmatrix} \quad (21a)$$

where

$$\begin{aligned} K_{ij}^{11} &= \int_{\Omega^e} [c_1 \frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} + c_2 \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial y} + c_3 \frac{\partial^2 \phi_i}{\partial x^2} \frac{\partial^2 \phi_j}{\partial x^2} + c_4 \frac{\partial^2 \phi_i}{\partial y^2} \frac{\partial^2 \phi_j}{\partial y^2} \\ &\quad + c_5 (\frac{\partial^2 \phi_i}{\partial x^2} \frac{\partial^2 \phi_j}{\partial y^2} + \frac{\partial^2 \phi_i}{\partial y^2} \frac{\partial^2 \phi_j}{\partial x^2}) + 4c_6 \frac{\partial^2 \phi_i}{\partial x \partial y} \frac{\partial^2 \phi_j}{\partial x \partial y}] dx dy \\ K_{iJ}^{12} &= \int_{\Omega^e} [c_1 \frac{\partial \phi_i}{\partial x} \psi_J + c_7 \frac{\partial^2 \phi_i}{\partial x^2} \frac{\partial \psi_J}{\partial x} + c_9 \frac{\partial^2 \phi_i}{\partial y^2} \frac{\partial \psi_J}{\partial y} \\ &\quad + 2c_{10} \frac{\partial^2 \phi_i}{\partial x \partial y} \frac{\partial \psi_J}{\partial y}] dx dy \end{aligned}$$

$$K_{IJ}^{13} = \int_{\Omega} e \left[c_2 \frac{\partial \phi_i}{\partial y} \psi_J + c_8 \frac{\partial^2 \phi_i}{\partial y^2} \frac{\partial \psi_J}{\partial y} + c_9 \frac{\partial^2 \phi_i}{\partial x^2} \frac{\partial \psi_J}{\partial y} + 2c_{10} \frac{\partial^2 \phi_i}{\partial x \partial y} \frac{\partial \psi_J}{\partial x} \right] dx dy$$

$$K_{IJ}^{22} = \int_{\Omega} e \left[c_1 \psi_I \psi_J + c_{11} \frac{\partial \psi_I}{\partial x} \frac{\partial \psi_J}{\partial x} + c_{14} \frac{\partial \psi_I}{\partial y} \frac{\partial \psi_J}{\partial y} \right] dx dy$$

$$K_{IJ}^{23} = \int_{\Omega} e \left[c_{12} \frac{\partial \psi_I}{\partial x} \frac{\partial \psi_J}{\partial y} + c_{14} \frac{\partial \psi_I}{\partial y} \frac{\partial \psi_J}{\partial x} \right] dx dy$$

$$K_{IJ}^{33} = \int_{\Omega} e \left[c_2 \psi_I \psi_J + c_{13} \frac{\partial \psi_I}{\partial y} \frac{\partial \psi_J}{\partial y} + c_{14} \frac{\partial \psi_I}{\partial x} \frac{\partial \psi_J}{\partial x} \right] dx dy \quad (21b)$$

4.2 Mixed Model

To relax the continuity requirements placed in the displacement model described above, here we consider a mixed formulation of the problem. In a mixed model, independent approximations of the displacements and stress resultants are used. A mixed variational formulation of these equations can be obtained by treating u , v , w , ψ_x , ψ_y , M_1 , M_2 , M_6 , P_1 , P_2 and P_6 as the primary variables. The governing equations for u , v , w , ψ_x and ψ_y are given in Eq. (8). The equations for the other six variables (M_1 , M_2 , M_6 , P_1 , P_2 and P_6) are provided by the laminate constitute equations:

$$\begin{aligned} \{N\} &= [A]\{\epsilon^0\} + [B]\{\kappa^0\} + [E]\{\kappa^2\} \\ \{M\} &= [B]\{\epsilon^0\} + [D]\{\kappa^0\} + [F]\{\kappa^2\} \\ \{P\} &= [E]\{\epsilon^0\} + [F]\{\kappa^0\} + [H]\{\kappa^2\} \end{aligned} \quad (22a)$$

$$\begin{Bmatrix} Q_2 \\ Q_1 \\ R_2 \\ R_1 \end{Bmatrix} = \begin{bmatrix} A_{44} & A_{45} & D_{44} & D_{45} \\ & A_{55} & D_{45} & D_{55} \\ \text{symm.} & & F_{44} & F_{45} \\ & & & F_{55} \end{bmatrix} \begin{Bmatrix} \epsilon_4^0 \\ \epsilon_5^0 \\ \kappa_4^2 \\ \kappa_5^2 \end{Bmatrix} \quad (22b)$$

where A_{ij} , B_{ij} , etc., are the plate stiffnesses, defined by

$$(A_{ij}, B_{ij}, D_{ij}, E_{ij}, F_{ij}, H_{ij}) \\ = \int_{-h/2}^{h/2} Q_{ij}(1, z, z^2, z^3, z^4, z^6) dz \quad (i, j=1, 2, 6) \quad (23a)$$

$$(A_{ij}, D_{ij}, F_{ij}) \\ = \int_{-h/2}^{h/2} Q_{ij}(1, z^2, z^4) dz \quad (i, j=4, 5) \quad (23b)$$

The mixed model is of the form

$$[K^e]\{\Delta^e\} = \{F^e\} \quad (24)$$

where $[K^e]$, $\{\Delta^e\}$ and $\{F^e\}$ are the generalized element stiffness matrix, element displacement vector, and element force vector, respectively.

The vector $\{\Delta^e\}$ denotes the set of nodal values,

$$\{\Delta^e\}^T = \{u_1, u_2, \dots, u_n, v_1, v_2, \dots, v_n, w_1, w_2, \dots, w_n, \\ \psi_x^1, \psi_x^2, \dots, \psi_x^n, \psi_y^1, \psi_y^2, \dots, \psi_y^n, M_1^1, M_1^2, \dots, M_1^n, \\ M_2^1, M_2^2, \dots, M_2^n, M_6^1, M_6^2, \dots, M_6^n, P_1^1, P_1^2, \dots, P_1^n, \\ P_2^1, P_2^2, \dots, P_2^n, P_6^1, P_6^2, \dots, P_6^n\} (e) \quad (25)$$

The displacement model results in eight degrees of freedom ($u, v, w, \frac{\partial w}{\partial x}, \frac{\partial w}{\partial y}, \frac{\partial^2 w}{\partial x \partial y}, \psi_x, \psi_y$) per node while the mixed model results in eleven degrees of freedom ($u, v, w, \psi_x, \psi_y, M_1, M_2, M_6, P_1, P_2, P_6$) per node. The displacement model with linear approximation of u, v, ψ_x and ψ_y and Hermite cubic interpolation of w results in stiffness matrix of order 32×32 . For linear interpolation of all variables, the mixed element results in stiffness matrix of order 44×44 . Thus, the mixed element is computationally expensive compared to the displacement model, although increased accuracy of the bending moments is expected in the mixed model.

Acknowledgements

The research reported herein is conducted under a grant from the Army Research Office (Grant No. DAAG 29-85-K-0007). The support is gratefully acknowledged.

6. REFERENCES

- [1] Basset, A. B., "On the extension and flexure of cylindrical and spherical thin elastic shells," Phil. Trans. Royal Soc., (London), Ser. A, Vol. 181, No. 6, pp. 433-480, 1890.
- [2] Hildebrand, F. B., Reissner, E. and Thomas, G. B., "Note on the foundations of the theory of small displacements of orthotropic shells," NACA Technical Note No. 1833, March 1949.
- [3] Hencky, H., "Über die berücksichtigung der schubverzerrung in ebenen platten," Ing. Arch., Vol. 16, 1947.
- [4] Mindlin, R. D., "Influence of rotatory inertia and shear on flexural motions of isotropic, elastic plates," J. Appl. Mech., Vol. 18, pp. A31, 1951.
- [5] Reddy, J. N., "A review of the literature on finite element modeling of laminated composite plates," The Shock and Vibration Digest, Vol. 17, No. 4, pp. 1-8, 1985.
- [6] Levinson, M., "An accurate theory of the statics and dynamics of elastic plates," Mech. Res. Communi., Vol. 7, p. 343, 1980.
- [7] Murthy, M. V. V., "An improved transverse shear deformation theory for laminated anisotropic plates," NASA Technical Paper 1903, November 1981.
- [8] Reddy, J. N., "A refined nonlinear theory of plates with transverse shear deformation," Int. J. Solids Struct., Vol. 20, No. 9/10, pp. 881-896, 1984.
- [9] Reddy, J. N., "A simple higher-order theory for laminated composite plates," J. Appl. Mech., Vol. 51, pp. 745-752, 1984.
- [10] Phan, N. D. and Reddy, J. N., "Analysis of laminated plates using a higher-order shear deformation theory," Int. J. Numer. Meth. Engng., to appear (1985).
- [11] Putcha, N. S. and Reddy, J. N., "A refined mixed shear flexible finite element for the nonlinear analysis of laminated plates," Computers and Structures, to appear.

- [12] Reddy, J. N., Energy and variational methods in applied mechanics, John Wiley, New York, 1984.
- [13] Reddy, J. N., Applied functional analysis and variational methods in engineering, McGraw-Hill, New York, 1986.
- [14] Reddy, J. N., An introduction to the finite element method, McGraw-Hill, New York, 1984.

THREE PHASE FLOW IN A POROUS MEDIUM AND THE CLASSIFICATION OF NON-STRICTLY HYPERBOLIC CONSERVATION LAWS

Michael Shearer
Department of Mathematics
North Carolina State University
Raleigh, North Carolina 27695

David G. Schaeffer
Department of Mathematics
Duke University
Durham, North Carolina 27706

1. INTRODUCTION. A system of conservation laws in one space dimension

$$U_t + F(U)_x = 0 \quad -\infty < x < \infty, \quad t > 0 \quad (1.1)$$

is strictly hyperbolic at U if $dF(U)$ has distinct real eigenvalues. Many hyperbolic systems of physical interest fail to be strictly hyperbolic at every point. In this paper, we summarize recent work on 2×2 hyperbolic systems that are strictly hyperbolic except at a single point called an umbilic point.

Equations with umbilic points arise as models of three phase flow in a porous medium. In this context, the Riemann initial value problem assumes especial importance, since it is central to numerical front tracking methods based on Glimm's scheme. The Riemann problem for (1.1) has jump initial data

$$U(x,0) = \begin{cases} U_L & \text{if } x < 0 \\ U_R & \text{if } x > 0 \end{cases} \quad (1.2)$$

Solutions of (1.1), (1.2) involve combinations of centered shock and rarefaction waves. Each wave in a specific characteristic family may be characterized by the state U_- on the left of the wave in (x,t) -space and the wave strength. For fixed U_- , the strength of the wave parameterizes a one-parameter family of states U_+ on the right of the wave. These one-parameter families (one for each characteristic family) are called wave curves. The umbilic point has a profound effect on the structure of wave curves. An overall goal of this research is to classify the various effects produced by the presence of such a point. The classification of 2×2 systems in [3] identifies four different regimes that we describe in §2. In §3 we illustrate the effect of the umbilic point by describing the solution of a prototype Riemann problem. Of particular note are two new types of shock wave violat-

ing the Lax entropy condition. These overcompressive and undercompressive shocks are associated with the presence of the umbilic point, which confuses the labelling of the characteristic speeds. Details of the results and properties described here, together with new techniques for studying Riemann problems, are discussed fully in [3,4].

2. CLASSIFICATION OF EQUATIONS WITH UMBILIC POINTS. To motivate the analysis, we begin by describing equations modelling three phase flow in a porous medium. Let u, v, w denote volume fractions of the phases, with corresponding relative permeabilities f, g, h . We assume the flow is one-dimensional, is not subject to external forces, and that the pressure in each of the phases is the same. We make the constitutive, or modelling, assumption that f, g, h depend only on u, v, w respectively: $f = f(u)$, etc. As usual, the conservation of momentum is approximated by Darcy's law, which for 1-d flows enables us to express the velocities in terms of the volume fractions. This puts the conservation of mass in the following form:

$$\left. \begin{aligned} u_t + \left(\frac{f(u)}{D} \right)_x &= 0 \\ v_t + \left(\frac{g(v)}{D} \right)_x &= 0, \end{aligned} \right\} \quad (2.1)$$

with $D = f(u) + g(v) + h(w)$ and $u + v + w = 1$. The physical range for u, v and $w = 1 - u - v$ is $[0, 1]$. Note that the corresponding equation for w is redundant and thus excluded from (2.1), and that f, g, h are taken to include the corresponding viscosities.

Here are properties of f, g, h that reflect experimental results for 2-phase flow:

$$(A) \quad f(0) = f'(0) = 0, \quad f''(u) > 0, \quad 0 \leq u \leq 1,$$

and similarly for g, h .

Let us return to terminology for general 2×2 systems in order to state a result for system (2.1):

$$U_t + F(U)_x = 0, \quad (2.2)$$

where $U = U(x, t) \in \mathbb{R}^2$, $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$. System (2.2) is hyperbolic if $dF(U)$ has real eigenvalues (characteristic speeds), which we label $\lambda_1(U) \leq \lambda_2(U)$. If $\lambda_1(U) = \lambda_2(U)$, we call U an umbilic point.

An umbilic point U^* is essential if the set of matrices

$\gamma = \{dF(U): U \text{ near } U^*\}$ is a smooth 2-dimensional manifold, in

which U^* is the only umbilic point, and $dF(U^*)$ is diagonal.

The term "essential" reflects the property that such umbilic points cannot be removed by perturbations of F .

Theorem 2.1. Under assumption (A), system (2.1) has a unique essential umbilic point in the triangle $0 < u, v, 1 - u - v < 1$.

We next turn to the structure of wave curves for system (2.2) near an essential umbilic point U^* . Consider the Taylor series for $F(U)$ about U^*

$$F(U) = F(U^*) + dF(U^*)(U - U^*) + Q(U - U^*) + \text{h.o.t.},$$

where $Q = \frac{1}{2} d^2 F(U^*)$ and h.o.t. indicates the remainder term.

Since $F(U^*)$ is constant and $dF(U^*)$ is a multiple of the identity (and may be removed by changing to a moving system of coordinates), the first term to affect wave curves is $Q(U - U^*)$. Assuming Q is nondegenerate in a sense specified in [3], the higher order terms do not change the qualitative properties of wave curves near the umbilic point. We shall ignore the higher order terms in what follows. Taking $U^* = 0$ without loss of generality, we are left with the system

$$U_t + Q(U)_x = 0 \quad -\infty < x < \infty, \quad t > 0 \quad (2.3)$$

Now $U = 0$ is automatically an umbilic point for (2.3), and is essential if $dQ(U)$ has distinct (real) eigenvalues for every $U \neq 0$.

Now linear changes of dependent variable U in (2.3) do not affect the characteristic speeds, or other features such as the admissible shock waves and the rarefaction waves. Accordingly, it is appropriate in a classification of equations (2.3) to call two equations equivalent if a linear change of variable converts one to the other. That is, quadratic maps $Q_k: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ($k = 1, 2$) are equivalent if there exists a (constant) matrix S such that $Q_1(SU) = SQ_2(U)$ for all $U \in \mathbb{R}^2$. If $Q(U) = dC(U)$ for a homogeneous cubic scalar $C: \mathbb{R}^3 \rightarrow \mathbb{R}$, then (2.3) is symmetric, and automatically hyperbolic. The following result says that up to equivalence, all equations (2.3) with an essential umbilic point are symmetric.

Theorem 2.2. If $U = 0$ is an essential umbilic point for (2.3), then Q is equivalent to dC , where

$$C(u, v) = au^3/3 + bu^2v + uv^2 \quad (2.4)$$

and $a \neq 1 + b^2$.

Note that (2.4) is not the most general cubic scalar. We have rotated coordinates to eliminate the v^3 term and scaled to let the coefficient of uv^2 be unity. Both of these transformations preserve equivalence. Theorem 2.2 reduces the study of equations (2.3) to an investigation of a 2-parameter family of equations:

$$U_t + dC(U)_x = 0 \quad (2.5)$$

Recall that shock waves for a 2×2 system are compressive if they satisfy the Lax entropy condition [1], which requires characteristics of one family to enter the shock from both sides and characteristics of the other family to pass through the shock. There are also what we call overcompressive shocks for which both families of characteristics enter the shock, and noncompressive shocks for which both families of characteristics pass through the shock. If we fix U_L , all possible shock waves with U_L on the left are identified by the set of U_R lying on a shock wave curve.

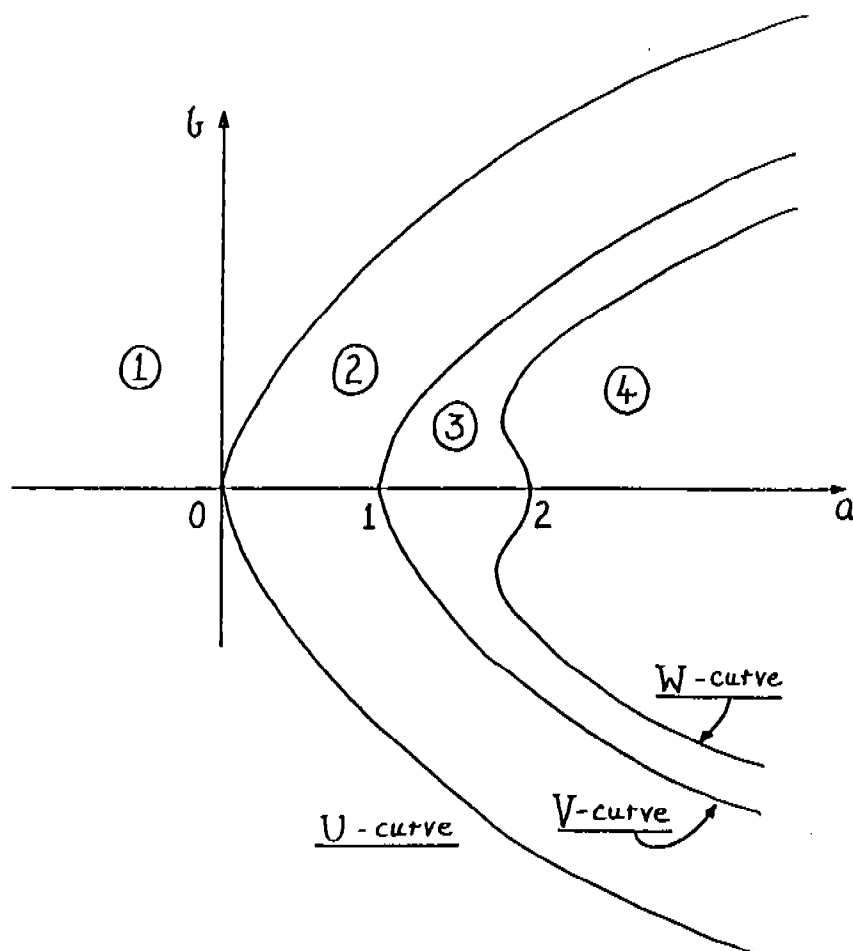
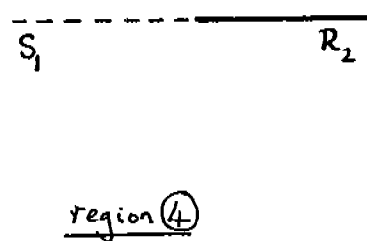
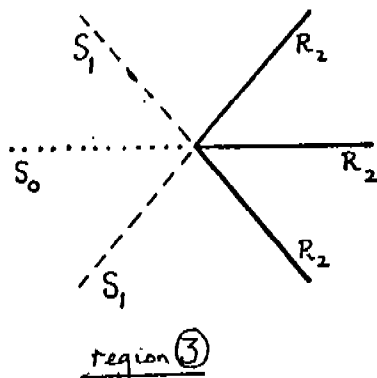
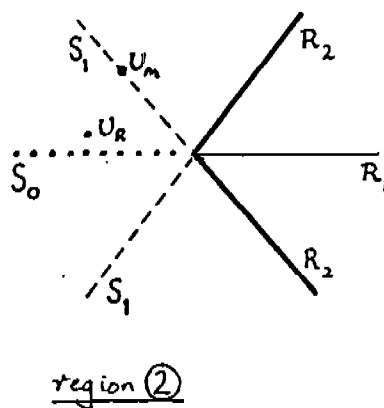
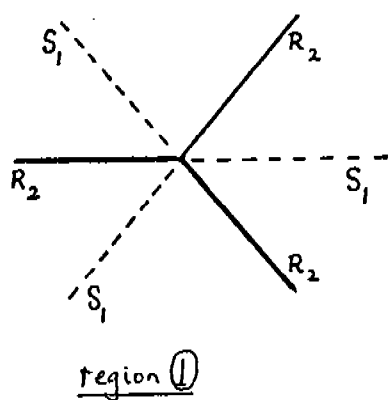


Figure 1. Regions in the (a, b) -plane.

These U_R satisfy the Rankine-Hugoniot condition and correspond to compressive and overcompressive shocks. (Certain noncompressive shocks also need to be selected in general. This is discussed briefly in §3.) We also have rarefaction curves which give the values of U through a centered rarefaction wave. These values lie on an integral curve of one of the two right eigenvectors of $dF(U)$. In Figure 2, we show these wave curves for (2.5) with $U_L = 0$ on the left of the wave. There are four regions in which the structure of the wave curves changes; these are shown in Figure 1.



S_1 : slow shocks

S_0 : overcompressive shocks

R_1 : slow rarefactions

R_2 : fast rarefactions

Figure 2. Wave curves originating at the umbilic point.

To understand the overcompressive shocks, consider the Riemann problem (1.2), (2.5) in region 2, when $U_L = 0$ and U_R is close to the curve S_0 representing overcompressive shocks in Figure 2. The solution involves two compressive shock waves of nearly equal speed separated by a state U_m on the curve S_1 . This is illustrated in Figure 3. Thus, an overcompressive shock wave is a superposition of fast and slow compressive shock waves travelling at the same speed.

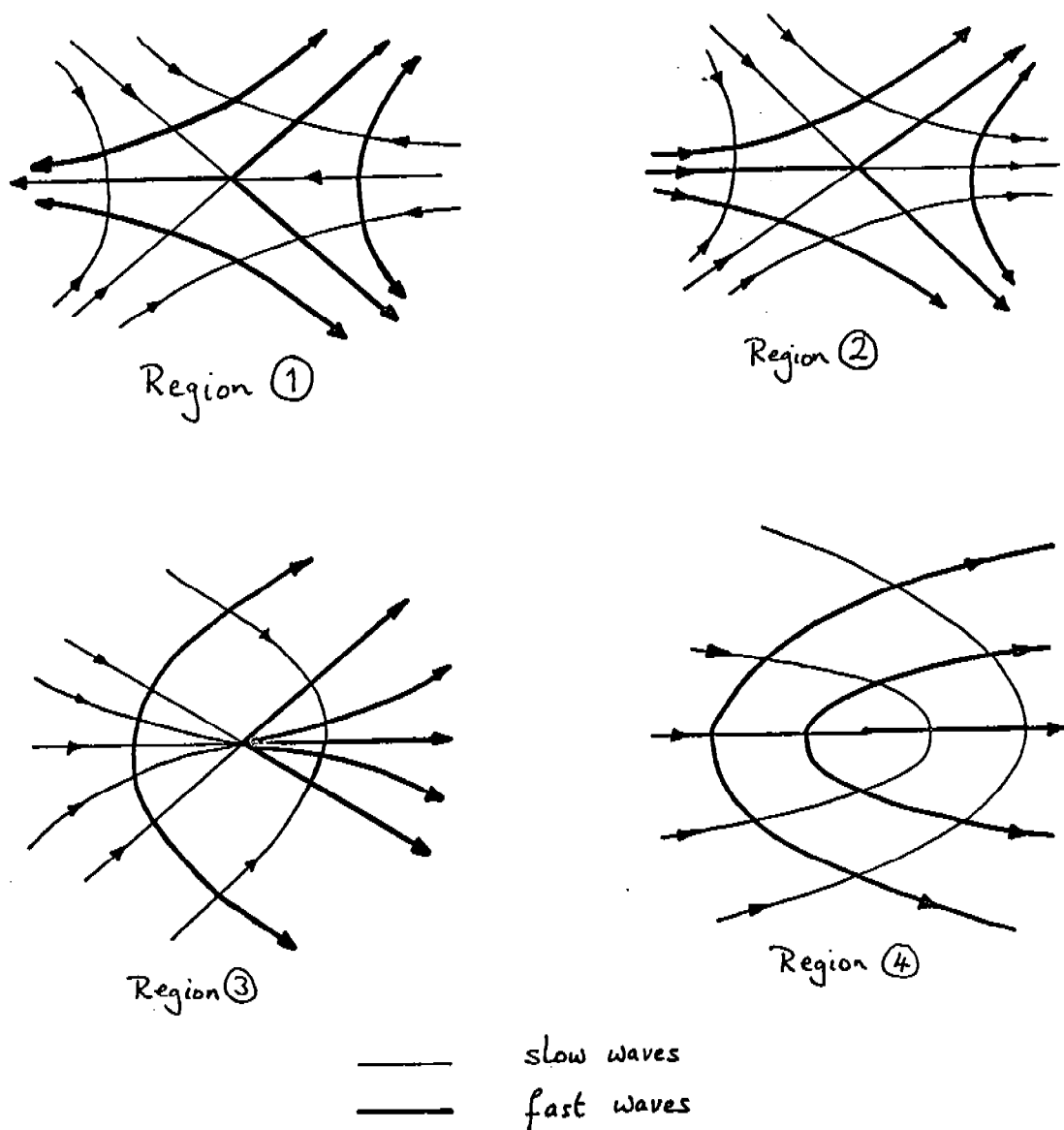


Figure 3. Rarefaction curves.

In Figure 4, we show the patterns of rarefaction curves for (2.5) in the four cases. Note that the arrows indicate the direction of increasing characteristic speed.

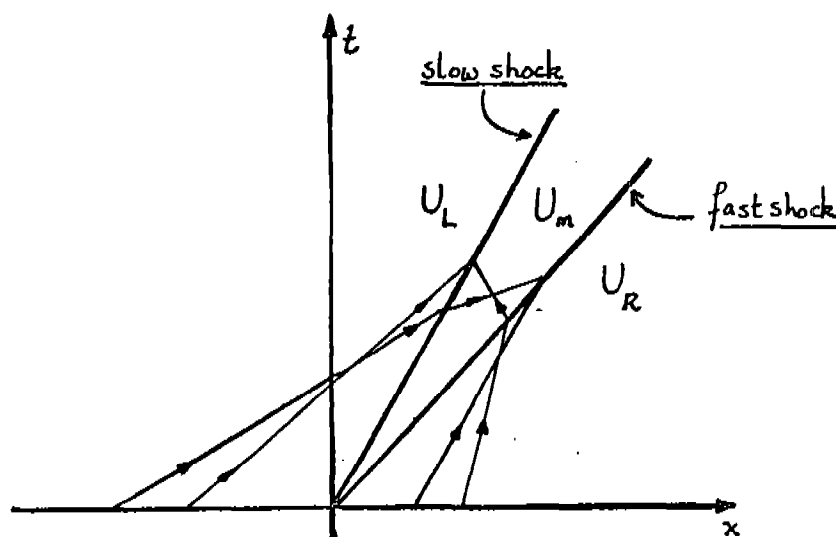


Figure 4. Solution of Riemann problem.
 $U_L = 0$, U_R near S_0 .

The following result shows how the 3-phase flow model (2.1) fits into the classification. By Theorem 2.1, system (2.1) has an umbilic point (u^*, v^*) . Let Q denote the quadratic terms of the Taylor expansion of the nonlinearity of (2.1) about (u^*, v^*) .

Theorem 2.3. Under assumption (A), the quadratic mapping Q for system (2.1) is equivalent to dC , for C given by (2.4) with $a < 1 + b^2$.

In other words, system (2.1) has properties near (u^*, v^*) corresponding only to quadratic nonlinearities in regions 1 and 2 of Figure 1.

3. SOLUTION OF A RIEMANN PROBLEM. In this section, we explain the solution of the Riemann problem

$$z_t - (\bar{z}^2)_x = 0 \quad (3.1)$$

$$z(x, 0) = \begin{cases} z_L & x < 0 \\ z_R & x > 0 \end{cases}, \quad (3.2)$$

where $z = u + iv$. Note that equation (3.1) has the symmetry property that $e^{2\pi i/3} z$ and \bar{z} are solutions whenever z is a solu-

tion. For each z_L, z_R , there is a combination of centered shock and rarefaction waves that join z_L to z_R . For fixed z_L , each combination of successively faster waves determines a value of z_R , and as the strengths of the waves change, so does the value of z_R . In Figure 5, each possible combination of shock waves (S) and rarefaction waves (R) is indicated and associated with a region.

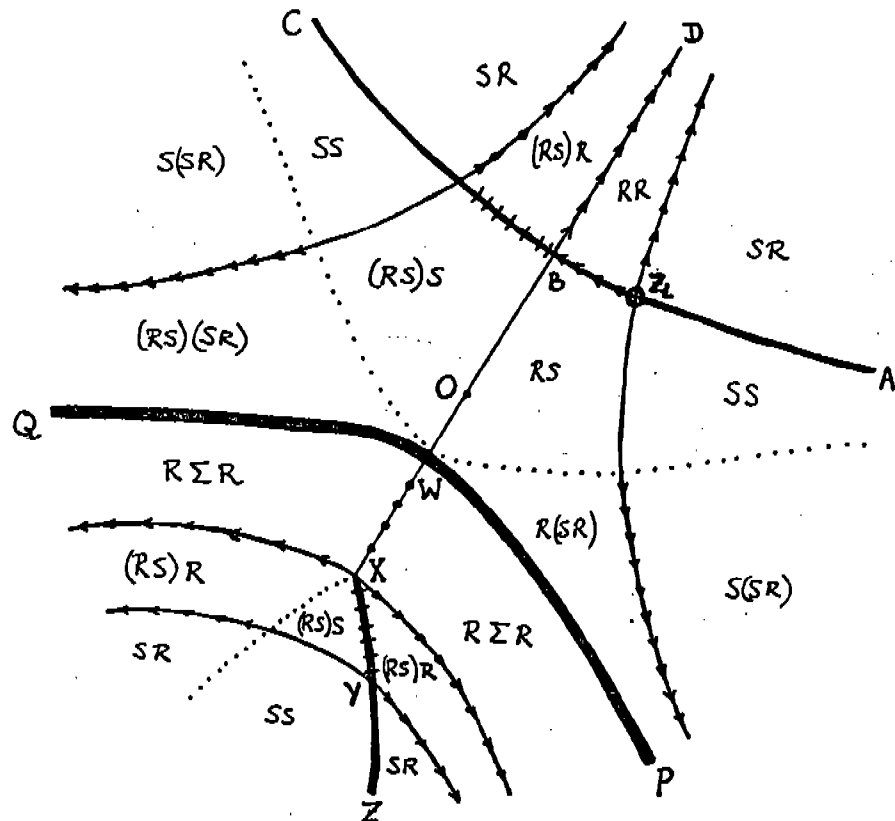


Figure 5. Solution of the Riemann problem (3.1), (3.2) for a typical z_L .

For example, if z_R lies in a region labelled (RS)S, then the solution of the Riemann problem involves a slow rarefaction-shock, and a fast shock wave. The label Σ indicates an undercompressive shock. These are noncompressive shocks that have viscous profiles associated with the Burgers' system

$$z_t - (\bar{z}^2)_x = \epsilon z_{xx} . \quad (3.3)$$

Now as z_L changes, the z_R regions get distorted. The regions have the shapes indicated in Figure 5 only for z_L between the lines $M_1^+ = \{z > 0\}$ and $M_2^- = \{z = \alpha e^{i\pi/3} : \alpha > 0\}$, which are lines of symmetry for equation (3.1). As z_L crosses M_2^- or M_1^+ , several regions coalesce and collapse, and then reform in a different arrangement that is easily obtained by symmetry.

One interesting feature of Figure 5 emphasizes the role of the umbilic point $z = 0$. The curve PQ separates two different constructions of solutions of the Riemann problem. If z_R lies above PQ, then the standard construction of Liu, for strictly hyperbolic systems, applies. This says that there is a point z_m on the curve ABC (representing slow waves) such that z_R lies on the curve through z_m representing fast waves. Below PQ, the construction is similar except that the slower waves (including the undercompressive shocks) are represented by the curve WXYZ, which does not include z_L . The "coordinate system" that was used above PQ is, roughly speaking, rotated through 90° below PQ.

It should be emphasized that the solution of (3.1), (3.2) has relied upon the symmetry property of (3.3) to distinguish the undercompressive shocks. Without the symmetry, or if the diffusion matrix fails to be a multiple of the identity, then characterizing shocks possessing viscous profiles is significantly more difficult.

Reference

1. P. D. Lax. Hyperbolic systems of conservation laws II. Comm. Pure Appl. Math. 10 (1957), 537-566.
2. T.-P. Liu. The Riemann problem for general 2×2 conservation laws. Trans. Amer. Math. Soc. 199 (1974), 89-112.
3. D. G. Schaeffer and M. Shearer. The classification of non-strictly hyperbolic conservation laws, with application to oil recovery. Submitted to Comm. Pure Appl. Math.
4. M. Shearer, D. G. Schaeffer, D. Marchesin and P. L. Paes-Leme. Solution of the Riemann problem for a prototype 2×2 system of non-strictly hyperbolic conservation laws. Submitted to Arch. Rat. Mech. Anal.

A GENERALIZED HEAT EQUATION: AN OVERVIEW

Siegfried H. Lehnigk
Research, Development, and Engineering Center
U.S. Army Missile Command
Redstone Arsenal, Alabama 35898-5248

ABSTRACT. The basic equation to be discussed is:

$$\frac{\partial}{\partial x} \left[A(x) \frac{\partial z}{\partial x} + D(x)z \right] - \frac{\partial z}{\partial t} = 0, \quad x > 0, \quad t > 0, \quad (*)$$

$$A(x) = \alpha x^{\lambda+1}, \quad D(x) = \alpha p x^{\lambda} + \beta x,$$

with parameters $\alpha > 0, \lambda < 1, p < 1, \beta \in \mathbb{R}$. The standard heat equation arises for $\lambda = -1, p = \beta = 0$. Crucial for the construction of solutions of (*) is a function $v^*(x, t; y)$ which is derived from the fundamental solution $v(x, t; y, s)$. $v^*(x, t; y)$ is the delta function initial condition solution of (*) with δ applied at $x = y = 0$, describing distribution processes from a completely concentrated initial state. $v^*(x, t; y)$ is the kernel for Poisson-Lebesgue and Poisson-Stieltjes transform solutions of (*). It also forms the basis for the construction of bi-orthogonal solution sequences of (*). Furthermore, it is the essential ingredient for the definition of a generalized Jacobi Theta function:

$$\theta(x, t) = v^*(x, t; 0) + 2 \sum_{n=1}^{\infty} v^*(x, t; y_n).$$

For fixed $t > 0$, the function $v^*(x, t; 0)$ may be interpreted as a probability density function. It contains as special cases a number of well known functions.

I. ORIGIN OF THE EQUATION

The equation

$$\frac{\partial}{\partial x} \left[A(x) \frac{\partial z}{\partial x} + D(x)z \right] - \frac{\partial z}{\partial t} = 0, \quad z = z(x, t), \quad (1.1)$$

was used as a (linear) one-dimensional autonomous model for various diffusion type processes in problems of heat conduction, diffusion of atoms into semiconductor materials, and decay of molecular systems from excited states.

For physical reasons, the requirement arose that equation (1.1) be of such a nature that it admits a conservative similarity solution in the open first quadrant of the (x, t) plane. A solution of this kind is, by definition, of the general form

$$z_0(x, t) = b^{-1}(t)f(\xi), \quad \xi = xb^{-1}(t), \quad (1.2)$$

with $b(0)=0$, $b(t) > 0$ and increasing as $t \uparrow \infty$, and $f(\xi)$ such that

$$\int_0^{\infty} z_0(x, t) dx \equiv 1.$$

Under appropriate circumstances, the function (1.2) will be the delta function initial condition solution of equation (1.1) with the delta function applied at the origin $(x, t)=(0, 0)$. It will then describe a distribution process in space and time from a completely concentrated initial state at $(0, 0)$.

Substitution of (1.2) into (1.1) leads to a condition on the coefficient functions $A(x)$ and $D(x)$ of the form

$$F(A, \frac{dA}{dx}, D, \frac{dD}{dx}) = 0 \quad (1.3)$$

and to a first order equation for $b(t)$ and a second order equation for $f(\xi)$. An additional requirement is needed to determine A, D, b , and f uniquely. It could be derived from a physical principle associated with the process to be modeled by equation (1.1) or from an assumption about the physical nature of the underlying diffusion process. It was decided that the diffusion coefficient $A(x)$ should satisfy a power law,

$$A(x) = \alpha x^{\lambda+1}, \quad \alpha > 0.$$

Relation (1.3) then leads to the drift coefficient

$$D(x) = \alpha p x^{\lambda} + \beta x, \quad \beta \in \mathbb{R},$$

and the functions $b(t)$ and $f(\xi)$ appearing in (1.2) take the specific forms

$$b(t) = \begin{cases} [\alpha(1-\lambda)^2 t]^{(1-\lambda)^{-1}}, & \beta = 0, \\ [\alpha(1-\lambda)\beta^{-1} (1 - \exp - (1-\lambda)\beta t)]^{(1-\lambda)^{-1}}, & \beta \neq 0, \end{cases} \quad (1.4)$$

$$f(\xi) = \frac{1-\lambda}{\Gamma(1+q)} \xi^{-p} \exp -\xi^{1-\lambda}, \quad \xi = xb^{-1}(t), \quad q = (\lambda-p)(1-\lambda)^{-1}. \quad (1.5)$$

It is clear that the parameter value $\lambda = 1$ must be avoided and that, therefore, $\lambda > 1$ or $\lambda < 1$. It turns out that $\lambda < 1$ leads to physically interesting situations. Because of the conservativeness requirement, it is necessary to restrict the parameter p to values $p < 1$. This then makes the compound parameter q appearing in (1.5) greater than -1 .

Special cases of the now specified equation (1.1) are the heat equation [1] for $\lambda = -1$, $p = \beta = 0$, and the Feller equation [2,3] for $\lambda = 0$.

II. A FUNDAMENTAL SOLUTION

The equation

$$\frac{\partial}{\partial x} [A(x) \frac{\partial z}{\partial x} + D(x)z] - \frac{\partial z}{\partial t} = 0, \quad (2.1a)$$

$$A(x) = \alpha x^{\lambda+1}, \quad D(x) = \alpha p x^{\lambda} + \beta x, \quad \alpha > 0, \lambda < 1, p < 1, \beta \in \mathbb{R} \quad (2.1b)$$

has a fundamental solution $v(x, t; y, s)$, $x > 0$, $t > 0$, $y > 0$, $s \geq 0$ [4]. This is to say that $v(x, t; y, s)$ for fixed y and s is a solution of (2.1) as a function of $x > 0$, $t > 0$, and for x and t fixed it is a solution of the corresponding adjoint equation for $y > 0$, $s \geq 0$.

A special solution of equation (2.1) can be obtained from the fundamental solution

$$\begin{aligned} v^*(x, t; y) &= (1-\lambda)y^{-(1+\lambda)} v(x, t; y, 0) \\ &= (1-\lambda)b^{-1}\xi^{-\frac{1}{2}(p+\lambda)} (e^{-\beta t\eta})^{\frac{1}{2}(p-\lambda)} \\ &\quad \times I_q(2\xi^{\frac{1}{2}(1-\lambda)}(e^{-\beta t\eta})^{\frac{1}{2}(1-\lambda)}) \exp(-\xi^{1-\lambda} - (e^{-\beta t\eta})^{1-\lambda}), \end{aligned} \quad (2.2)$$

$\xi = xb^{-1}$, $\eta = yb^{-1}$, $q = (\lambda - p)(1-\lambda)^{-1}$, $b = b(t)$ as given by (1.4), I_q = modified Bessel function of the first kind (of imaginary argument)

III. OTHER SPECIAL SOLUTIONS

The function $v^*(x, t; y)$ defined by (2.2) plays a crucial role in the construction of other special solutions of the equation (2.1).

Boundary condition solutions of (2.1) for prescribed behavior along the t -axis have been established in [5]. They will not be discussed further in this overview.

Initial condition solutions can be defined as Poisson-Lebesgue transforms

$$z(x,t) = \int_0^{\infty} v^*(x,t;y)g(y)dy, \quad (3.1)$$

provided that $g(x)$ is Lebesgue summable over every compact interval of $0 < x < \infty$. This class of solutions has the property that

$$z(x,0+) = g(x) \text{ almost everywhere,}$$

which justifies their designation as initial condition solutions.

Another important class of solutions can be defined as Poisson-Stieltjes transforms

$$z(x,t) = \int_0^{\infty} v^*(x,t;y)dh(y),$$

provided that $h(x)$ is of bounded variation on every compact interval of $0 < x < \infty$. A special case arises if one takes for $h(x)$ the Heaviside unit step function at $x = y > 0$. This situation identifies $v^*(x,t;y)$ as the delta function initial condition solution of equation (2.1) with the delta function applied at $x = y > 0$, $t = 0$. Letting $y \downarrow 0$ one obtains the particular solution

$$v^*(x,t;0+) = z_0(x,t) = \frac{1-\lambda}{\Gamma(1+q)} b^{-1}(t) \xi^{-p} \exp -\xi^{1-\lambda}. \quad (3.2)$$

Thus, the similarity solution (1.2) with b and f specified by (1.4) and (1.5), respectively, is being recovered and now identified as the delta function initial condition solution with the delta function applied at the origin.

For details on this section, the reader is referred to [4].

IV. BIORTHOGONAL SEQUENCES OF SOLUTIONS

Using the Poisson-Lebesgue transform (3.1) with $g(x) = x^{n(1-\lambda) - p}$
 $= \exp [(n(1-\lambda)-p)\log x]$ with $\log x$ real for $x > 0$, one can
define the sequence of initial condition solutions

$$v_n(x, t) = \int_0^\infty v^*(x, t; y) y^{n(1-\lambda) - p} dy \quad (n=0, 1, 2, \dots).$$

Furthermore, expanding $v^*(x, t; u^{(1-\lambda)^{-1}})$ into a power series
about $u = 0$, one obtains

$$v^*(x, t; u^{(1-\lambda)^{-1}}) = \sum_{n=0}^{\infty} \frac{1}{n!} w_n(x, t) u^n$$

which converges everywhere in the (finite) u -plane and, thus, is
an entire function of the complex variable u . It can be verified
directly that each of the coefficient functions $w_n(x, t)$ is a

solution of the equation (2.1) for $x > 0$, $t > 0$, although none of
them is an initial condition solution in the sense of (3.1).

The sequences $\{v_n(x, t)\}$ and $\{w_n(x, t)\}$ have the property that

$$\int_0^\infty x^p w_m(x, t) v_n(e^{\beta t} x, -t) dx = \begin{cases} 0 & \text{if } m \neq n, \\ n! \exp -[1+n(1-\lambda)]\beta t & \text{if } m = n. \end{cases}$$

This fact allows the expansion of certain types of solutions
 $z(x, t)$ of equation (2.1) into series in terms of the functions
 $v_n(x, t)$ for example.

A typical theorem says roughly that if a solution $z(x, t)$ of (2.1)
is of a certain structure, then $z(x, t) = \sum_{n=0}^{\infty} c_n v_n(x, t)$ and vice
versa. Details on this subject may be found in [6].

V. A GENERALIZED JACOBI THETA FUNCTION

This topic has been developed in [7] and the reader is referred to this paper for details. In terms of the special solution (2.2) of the equation (2.1) the function

$$\theta(x,t) = v^*(x,t;0) + 2 \sum_{n=1}^{\infty} v^*(x,t;y_n), x \in \mathbb{C}, t \in \mathbb{C}, \quad (5.1)$$

is being defined for a suitable sequence $\{y_n\}$, $0 < y_n < y_{n+1}$, $y_n \uparrow \infty$ as $n \uparrow \infty$, as a function of the complex variables x and t .

A special case of (5.1) arises for $\lambda = -1$, $p = \beta = 0$ (heat equation), and $y_n = n\pi$,

$$\theta(x,t) = \frac{2}{\pi} \vartheta_3(x, e^{-4\alpha t}), x \in \mathbb{C}, t \in \mathbb{C}, \operatorname{Re} t > 0, \quad (5.2)$$

ϑ_3 being one of the Jacobi Theta functions [8], Chap. XXI.

The analytic properties of the function $\theta(x,t)$ which, because of the relation (5.2), is being designated a generalized Jacobi Theta function, have been discussed in [7]. Typical results are as follows:

Theorem. For fixed $t \in \mathbb{C}$ such that $\operatorname{Re} b^{-(1-\lambda)}(-t) < 0$ and for

suitable $\{y_n\}$, $\theta(x,t)$ is a holomorphic function of

$x \in \{x \in \mathbb{C} : x \neq 0, |\arg x| < \pi\}$.

Termwise, differentiation is justified.

Remarks.

a. The set of points t such that $\operatorname{Re} b^{-(1-\lambda)}(-t) < 0$ is not empty. It contains all points t with $\operatorname{Re} t > 0$, $\operatorname{Im} t = 0$.

b. In special cases, the holomorphic nature of $\theta(x,t)$ extends to all of \mathbb{C} , for example in the case specified by (5.2).

Theorem. For fixed $x \in \{x \in \mathbb{C} : x \neq 0, |\arg x| < \pi\}$ and for

suitable $\{y_n\}$ $\theta(x,t)$ is a holomorphic function of $t \in \{t \in \mathbb{C} : \operatorname{Re} t > 0\}$ if $\beta = 0$, of $t \in H$ if $\beta < 0$, and of $t \in K$ if $\beta > 0$.

Termwise differentiation is justified.

Remarks.

a. Because of periodicity for $\beta \neq 0$, it is sufficient to consider only the behavior of $\theta(x,t)$ in the periodicity strip $\operatorname{Re} \sigma \in (-\infty, \infty)$, $\operatorname{Im} t = \omega \in$

$(-\frac{\pi}{(1-\lambda)|\beta|}, \frac{\pi}{(1-\lambda)|\beta|}]$. The sets H and K of the theorem are the unshaded domains shown in Figures 1 and 2, respectively.

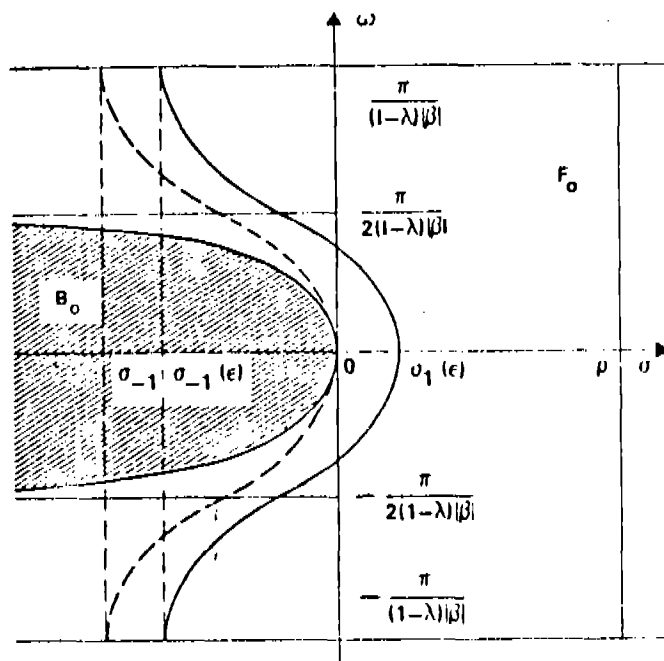


Fig. 1

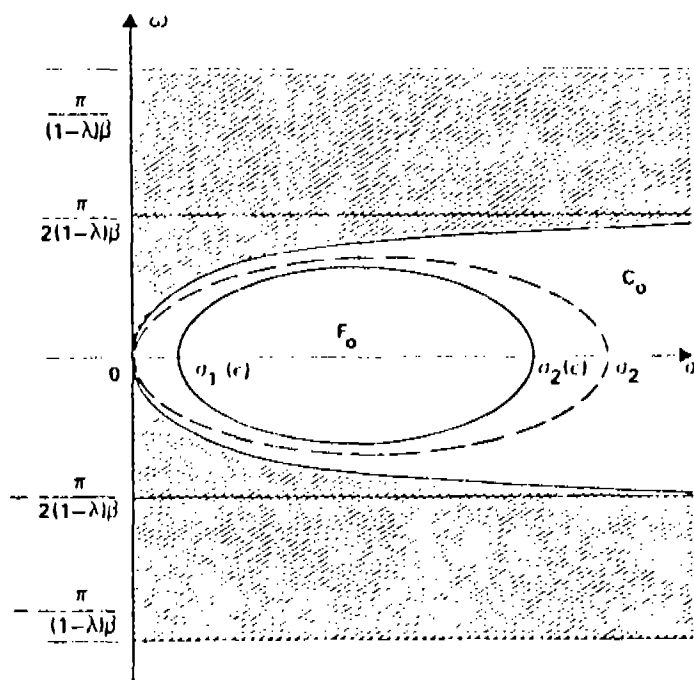


Fig. 2

b. The different behavior of $\theta(x, t)$ as a function of t relative to the values of the drift parameter β is not surprising. First of all, the different structure of the function $b(t)$ for $\beta = 0$ and for $\beta \neq 0$ as shown in (1.4) should be observed. Secondly, it is important to note that $\beta > 0$ as compared with $\beta < 0$ results in a drastic change in the behavior of the distribution process

$$v^*(x, t; y), y \geq 0. \text{ If } \beta > 0, b(t) \uparrow [\alpha(1-\lambda)\beta^{-1}]^{(1-\lambda)^{-1}} > 0$$

as $t \uparrow \infty$, which is to say that the distribution process approaches a steady state as $t \uparrow \infty$. This is in contrast to the case $\beta \leq 0$ in which $b(t) \uparrow \infty$ as $t \uparrow \infty$.

Theorem. $\theta(x, t)$ is a particular solution of equation (2.1).

VI. PROBABILISTIC CONSIDERATIONS

The real valued function of the real variable x

$$F(x) = \begin{cases} \frac{1}{\Gamma(1+q)} \gamma(1+q, \xi^{1-\lambda}), & x > 0, \xi = xb^{-1}, b > 0, \lambda < 1, p < 1, \\ & q = (\lambda-p)(1-\lambda)^{-1} > -1, \\ 0, & x \leq 0, \end{cases}$$

where $\gamma(a, y)$ is the incomplete Gamma function, may be designated a probability distribution function. The nondecreasing function $F(x)$ has the properties $F(x) > 0$, $F(x) \downarrow 0$ as $x \downarrow -\infty$, $F(x) \uparrow 1$ as $x \uparrow \infty$. Consequently, the real valued function of the real variable x

$$\phi(x) = \begin{cases} \frac{dF(x)}{dx} = \frac{1-\lambda}{\Gamma(1+q)} b^{-1} \xi^{-p} \exp -\xi^{1-\lambda}, & x > 0 \\ 0, & x \leq 0, \end{cases} \quad (6.1)$$

represents a probability density function. One recognizes immediately that $\theta(x)$ coincides with the delta function initial condition solution (3.2) of the equation (2.1) if t is being considered as a (positive) parameter.

The probability density function $\phi(x)$ given by (6.1) contains the following well-known functions as special cases:

a. the Gauss (normal) pdf

$$\frac{2}{\sqrt{\pi}} b^{-1} e^{-(x/b)^2}$$

for $\lambda = -1$, $p=0$.

b. the Weibull pdf

$$(1-\lambda)b^{-1}(x/b)^{-\lambda} e^{-(x/b)^{1-\lambda}}$$

for $p = \lambda$.

c. the negative exponential pdf

$$b^{-1} e^{-x/b}$$

for $\lambda = p = 0$ (special case of Weibull).

d. the Gamma pdf

$$\frac{1}{\Gamma(1-p)} b^{-1}(x/b)^{-p} e^{-x/b}$$

for $\lambda = 0$.

e. the Rayleigh pdf

$$2b^{-1}(x/b) e^{-(x/b)^2}$$

for $\lambda = p = -1$ (special case of Weibull).

f. the Maxwell pdf (Maxwell distribution for the absolute value of velocity)

$$\frac{4}{\sqrt{\pi}} b^{-1}(x/b)^2 e^{-(x/b)^2}$$

for $\lambda = -1$, $p = -2$.

g. the Wien pdf (Wien distribution for frequency, limiting case of Planck's distribution for high frequencies)

$$\frac{1}{6} b^{-1} \left(\frac{x}{b}\right)^3 e^{-x/b}$$

for $\lambda = 0$, $p = -3$.

It is useful to investigate the characteristic function associated with a probability density function. For the function (6.1) it is defined as

$$\Phi(s) = \frac{1-\lambda}{\Gamma(1+q)} b^{-1} \int_0^{\infty} \xi^{-p} e^{-\xi^{1-\lambda}} e^{s\xi} d\xi, \quad s = \sigma + i\omega,$$

or, upon the substitution $x = b\xi$,

$$\Phi(s) = \frac{1-\lambda}{\Gamma(1+q)} \int_0^{\infty} \xi^{-p} e^{-\xi^{1-\lambda} + bs\xi} d\xi. \quad (6.2)$$

The powers of ξ are, of course, defined in terms of $\log \xi$ with the principal value for $\log \xi$ for $\xi > 0$.

The transformation (6.2) defines $\Phi(s)$ as a holomorphic function

for $\operatorname{Re} s < 0$ if $0 < \lambda < 1$, for $\operatorname{Re} s < b^{-1}$ if $\lambda = 0$,

and for $s \in \mathbb{C}$ if $\lambda < 0$, i.e., for $\lambda < 0$, $\Phi(s)$ is an entire function of order $\rho = \lambda^{-1}(\lambda - 1)$, $1 < \rho < \infty$.

One has the power series expansion

$$\Phi(s) = \sum_{n=0}^{\infty} \frac{\alpha_n}{n!} s^n \quad (6.3)$$

with

$$\alpha_n = b^n \frac{\Gamma(1+q+n(1-\lambda))^{-1}}{\Gamma(1+q)} \quad (n=0,1,2,\dots)$$

being the moments of the probability density function (6.1). They exist for all values of $\lambda < 1$. However, for $0 < \lambda < 1$, the series evidently does not have a positive radius of convergence.

For $\lambda = 0$, the series (6.3) has the convergence radius $b^{-1} > 0$ and it represents the function

$$\Phi(s) = (1-bs)^{p-1} \quad (p < 1)$$

for $|s| < b^{-1}$. Clearly, for $\lambda < 0$, the convergence radius is $+\infty$.

In applications, it is useful to identify the parameter b appearing in (6.1) with the function $b(t)$ as defined by (1.4) to have the full set of parameters $\alpha > 0, \beta \in \mathbb{R}, \lambda < 1, p < 1, t < 0$ available.

REFERENCES

- [1] D. V. Widder, The Heat Equation, Academic Press, New York (1975).
- [2] W. Feller, Two singular diffusion problems, Ann. Math. 54, 173-182 (1951).
- [3] W. Feller, The parabolic differential equations and the associated semi-groups of transformations, Ann. Math. 55, 468-519 (1952).
- [4] S. H. Lehnigk, Initial condition solutions of the generalized Feller equation, J. Appl. Math. Phys. (ZAMP) 29, 273-294 (1978).
- [5] S. H. Lehnigk, Boundary condition solutions of the generalized Feller equation, J. Math. Phys. 19, 1267-1275 (1978).
- [6] S. H. Lehnigk, Biorthogonal sequences of solutions of the generalized Feller equation, Math. Meth. in the Appl. Sci. 4, 317-353 (1982).
- [7] S. H. Lehnigk, A generalized Jacobi theta function, Math. Meth. in the Appl. Sci. 6, 327-344 (1984).

ON THE NUMERICAL SOLUTION OF A STOCHASTIC
OPTIMAL CORRECTION PROBLEM*

P. L. Chow and J.L. Menaldi

Department of Mathematics, Wayne State University
Detroit, Michigan 48202

ABSTRACT The numerical solution to an optimal correction problem for a damped random linear oscillator is studied. A numerical algorithm for the discretized system of the governing variational inequalities will be given. To initiate the computation, we adopt a numerical scheme for the deterministic version of the problem. This will be followed by an algorithm based on a discrete maximum principle to ensure the convergence of the iteration process.

I. INTRODUCTION. In a previous paper [1], we consider the control problem for the damped linear oscillator excited by a random noise

$$\begin{cases} \ddot{x} + p\dot{x} + q^2x = r\dot{w}_t + \dot{v}_t, & 0 < t \leq T, \\ x(0) = x_0, & \dot{x}(0) = y_0 \end{cases} \quad (1)$$

where p, q are the damping and spring constants, and x_0, y_0 denote the initial position and velocity, respectively; r is the intensity of the white-noise \dot{w}_t ; v_t the control momentum at t , and T the horizon. Setting $y = \dot{x}$, Equation (1) may be rewritten in the Ito-differential form by a change of time scale:

*This work has been supported in part by the ARO Contract DAAG29-83-K-0014.

$$\begin{cases} dx_s = y_s ds, \\ dy_s = -(py_s + q x_s) ds + r dw_s + dv_s, \\ dt_s = ds, \end{cases} \quad (2)$$

$$x_0 = x, \quad y_0 = y, \quad t_0 = t.$$

For the admissible set V_{ad} of control v_t , we take the set of all processes of bounded variation, which depend on the Wiener process w_t in a nice way. Also, for simplicity, we assume the average cost function being of the special form

$$J_v(x, y, t) = E\{f(x_T, y_T) + c|v_T|\} \quad (3)$$

where f is smooth and of, at most, polynomial growth; $c > 0$ is a constant, and $|v_t|$ is the sum of the positive variation v_t^+ and the negative variation v_t^- of v_t . If we denote by u the minimum cost function

$$u(x, y, t) = \inf_v J_v(x, y, t), \quad (4)$$

then we can show that it satisfies the variational inequalities:

$$\begin{cases} (a) \quad \frac{\partial u}{\partial t} + Lu \geq 0, \\ (b) \quad \left| \frac{\partial u}{\partial y} \right| \leq c, \\ (c) \quad \left(\frac{\partial u}{\partial t} + Lu \right) \left(\left| \frac{\partial u}{\partial y} \right| - c \right) = 0, \quad u(x, y, T) = f(x, y). \end{cases} \quad (5)$$

where $0 < t < T$, $-\infty < x, y < \infty$, and

$$Lu = \frac{1}{2} r^2 \frac{\partial^2 u}{\partial y^2} - (2py + q x) \frac{\partial u}{\partial y} + y \frac{\partial u}{\partial x}. \quad (6)$$

To solve the problem numerically, we must replace the unbounded domain in the x - y - t space by a rectangular box $Q_t = \{(x, y, t) \text{ in } \mathbb{R}^3 : |x| \leq \ell_1, |y| \leq \ell_2, 0 \leq t \leq T\}$. Then we introduce a finite-difference approximation

to the differential inequalities in (5) and some appropriate boundary conditions on the boundary surface of Q_T . To this end we let $\Delta x = \ell_1/M$, $\Delta y = \ell_2/M$ and $\Delta t = T/N$ for some fixed integers $M, N > 0$. Let $Q_{M,N}$ denote the set of mesh points in Q_T , that is

$$Q_{M,N} = \{(x_i, y_j, t_n) : x_i = i\Delta x, y_j = j\Delta y, \\ t_n = n\Delta t; i, j = 0, \pm 1, \dots, \pm M, n = 0, 1, \dots, N\}.$$

The pivotal value of u at a mesh point $P(x_i, y_j, t_n)$ is given by

$$u_{i,j}^n = u(P) = u(x_i, y_j, t_n).$$

In what follows we shall present two numerical procedures for solving the problem (5) corresponding to two different finite-difference schemes for the variational inequalities. The first procedure is based on a deterministic version of the problem, while the second procedure deals with the full problem directly. For convenience, they will be called the first-order and second-order methods, respectively. The former method is simple and explicit, but less accurate. The latter is an implicit scheme whose convergence can be proven by means of a maximum principle. It seems that, by using the first-order procedure to obtain an initial approximation, the rate of convergence by the second-order procedure can be increased considerably.

We remark that the analysis of the control of the system (1) follows closely the techniques introduced in our earlier work [27], [3]. Some of the numerical techniques has been adopted from the work of Gronzalez and Rofman [4].

II. A FIRST-ORDER NUMERICAL METHOD. To avoid unnecessary algebraic complications, we set $p = 0$ in Eq. (1). Then we introduce a finite-difference approximation to the inequality (5.a) as follows:

$$\begin{aligned}
& \left\{ \frac{1}{\Delta t} (u_{i,j}^{n+1} - u_{i,j}^n) + \frac{r^2}{2(\Delta y)^2} (u_{i,j+1}^{n+1} - 2u_{i,j}^{n+1} + u_{i,j-1}^{n+1}) \right. \\
& + q^2 |i| \frac{\Delta x}{\Delta y} [h_i (u_{i,j-1}^n - u_{i,j}^n) + h'_i (u_{i,j+1}^n - u_{i,j}^n)] \\
& \left. + |j| \frac{\Delta y}{\Delta x} [h_j (u_{i+1,j}^n - u_{i,j}^n) + h'_j (u_{i-1,j}^n - u_{i,j}^n)] \right\} \geq 0,
\end{aligned} \tag{7}$$

$$\text{where } h_i = \begin{cases} 1, & \text{if } i \geq 0 \\ 0, & \text{if } i < 0, \text{ and } h'_i = (1-h_i). \end{cases}$$

By using the forward and the backward differences, the inequality (5.b) yields:

$$\begin{aligned}
u_{i,j}^n - u_{i,j-1}^n & \leq c \Delta y, \\
u_{i,j}^n - u_{i,j+1}^n & \leq c \Delta y.
\end{aligned} \tag{8}$$

The expressions (7) and (8) can be rewritten as:

$$\begin{aligned}
u_{i,j}^n & \leq a_{ij} (h_i u_{i,j-1}^n + h'_i u_{i,j+1}^n) \\
& + b_{ij} (h_i u_{i+1,j}^n + h'_i u_{i-1,j}^n) \\
& + c_{ij} (u_{i,j+1}^{n+1} - 2u_{i,j}^{n+1} + u_{i,j-1}^{n+1}),
\end{aligned} \tag{9}$$

$$\begin{cases} u_{i,j}^n \leq u_{i,j+1}^n + c \Delta y, \\ u_{i,j}^n \leq u_{i,j-1}^n + c \Delta y, \end{cases} \tag{10}$$

$$\begin{cases} u_{i,j}^n \leq u_{i,j+1}^n + c \Delta y, \\ u_{i,j}^n \leq u_{i,j-1}^n + c \Delta y, \end{cases} \tag{11}$$

where

$$a_{ij} = \frac{q^2 |i| \Delta x}{d_{ij} \Delta y}, \quad b_{ij} = \frac{|j| \Delta y}{d_{ij} \Delta x}, \quad c_{ij} = \frac{r^2}{2d_{ij} (\Delta y)},$$

$$d_{ij} = \left(\frac{1}{\Delta t} + q^2 |i| \frac{\Delta x}{\Delta y} + |j| \frac{\Delta y}{\Delta x} \right).$$

Let A_1 , A_2 , A_3 denote the right-hand sides of (9), (10) and (11), respectively.

Then the system of variational inequalities become

$$\left\{ \begin{array}{l} u_{i,j}^n \leq A_k u_{i,j}^n, \quad k=1,2,3, \\ \max_{1 \leq k \leq 3} \{ (u_{i,j}^n - A_k u_{i,j}^n) \} = 0, \quad n=0,1,\dots,N; \quad i,j=0,\pm 1,\dots,\pm M, \\ u_{i,j}^N = f_{i,j}, \\ u_{i,j} = g_{i,j}, \quad i=\pm M \text{ or } j=\pm M. \end{array} \right. \quad (12)$$

Here the last condition is an additional boundary condition needed when we replace the unbounded domain by a rectangular box. The first numerical algorithm is given as follows:

- 1.) Set $n = N$ and $u_{i,j}^N = f_{i,j}$, for $i,j=0,\pm 1,\dots,\pm N$.
- 2.) For $(i = -M; j=0,-1,\dots,-M),$
 $(i = M; j=0,1,\dots,M),$
 $(i = 0, -1,\dots,-M; j = M),$
 set $u_{i,j}^n = g_{i,j}$, with $n=0,\dots,N-1$.
- 3.) For $n=N-1$, calculate $u_{i,j}^n$ by the following steps:
 - (a) Set $u_{i,j}^n = u_{i,j}^{n-1}$ for $(i=0,1,\dots,M;j=0)$. From (12), compute

$$u_{i,j}^n = \min_{1 \leq k \leq 3} \{ A_k u_{i,j}^n \} \text{ for } j=0,1,\dots,M \text{ and } i=M,M-1,\dots,0.$$
 - (b) Knowing $u_{i,j}^n$ for $(i=0; j=0,\dots,M)$ and $(i=0,-1,\dots,-M; j=M)$, compute $u_{i,j}^n$ as before for $i=-1,\dots,-M$ and $j=0,-1,\dots,-M$.
 - (c) Knowing $u_{i,j}^n$ for $(i=0,-1,\dots,-M; j=0)$ and $(i=-M; j=0,-1,\dots,-M)$, find $u_{i,j}^n$ for $j=0,-1,\dots,-M$ and $i=-M, -M+1,\dots,0$.
 - (d) Knowing $u_{i,j}^n$ for $(i=0; j=0,-1,\dots,-M)$ and $(i=0,1,\dots,M; j=-M)$, find $u_{i,j}^n$ for $i=0,\dots,M$ and $j=-M, -M+1,\dots,0$.
- 4.) Check if the computed values of $u_{i,j}^n$ for $(i=0,1,\dots,M; j=0)$ agree with that of the initial guess. If yes, go to 6.). Otherwise, go to 5.).

- 5.) Adjust the initial values, say, by taking the mean of the guessed and the computed values and repeat the steps (3.a) - (3.d) until they agree with a prescribed error. Then go to 6.).
- 6.) Replace n by $(n-1)$ in 3.). Repeat the steps 3.) to 6.) and stop after $n=0$.

The above procedure can be schematically shown in the following Figure 1.

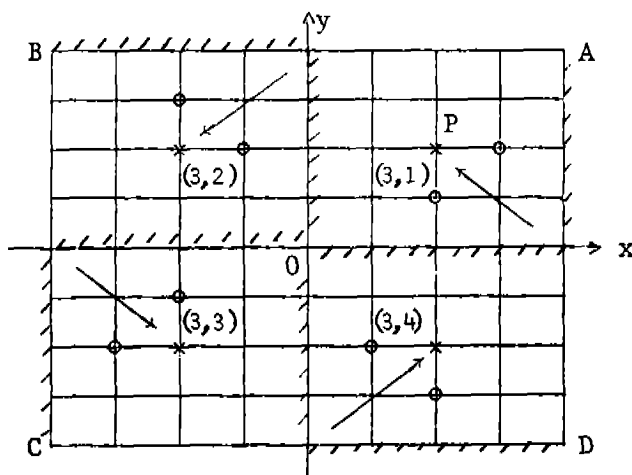


Figure 1

A First-Order Numerical Procedure

The above procedure is explicit in the sense that one can march backward in time. However, we do not expect a fast convergence of the scheme, if it converges at all. It is possible to devise a modified scheme to improve the accuracy of the method. But, since this procedure will be used only for the initial computations, we will not do so. For the major part of computations, the next algorithm, will be adopted to ensure a proper convergence.

III. AN ITERATIVE METHOD BASED ON MAXIMUM PRINCIPLE. In the left-hand side of (7), if we replace the superscript $n+1$ by n for each of three terms inside the second parenthesis, we would obtain the following finite-difference approximation to (5.1):

$$\begin{aligned} u_{i,j} \leq & \alpha_1 u_{i,j}^{n+1} + \alpha_2 (u_{i,j+1}^n + u_{i,j-1}^n) \\ & + \alpha_3 (h_i u_{i,j-1}^n + h'_i u_{i,j+1}^n) \\ & + \alpha_4 (h_j u_{i-1,j}^n + h'_j u_{i+1,j}^n), \end{aligned} \quad (13)$$

where

$$\begin{aligned} \alpha_1 &= \frac{1}{\alpha_0 \Delta t}, \quad \alpha_2 = \frac{r^2}{2\alpha_0 (\Delta y)}, \quad \alpha_3 = \frac{q^2 |i| \Delta x}{\alpha_0 \Delta y}, \quad \alpha_4 = \frac{|j| \Delta y}{\alpha_0 \Delta x}, \\ \alpha_0 &= \frac{1}{\Delta t} + \frac{r^2}{2(\Delta y)^2} + q^2 \frac{|i| \Delta x}{\Delta y} + \frac{|j| \Delta y}{\Delta x}. \end{aligned} \quad (14)$$

Here we note that α_k 's are positive with

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1, \quad (15)$$

and they depend on (i,j) .

For convenience, we regard the triple array $u_{i,j}^n$ as a vector u (or a tensor). Let

$$u = [u_{i,j}^n]^{(2M+1) \times (2M+1) \times (N+1)}$$

be a vector in $(2M+1)^2 \times (N+1)$ -dimensional space E with the maximum norm

$$\|u\| = \max_{i,j,n} |u_{i,j}^n|. \quad (16)$$

Let us define a linear operators M_1 , M_2 and M_3 on E such that

$$\left\{ \begin{array}{l} (M_1 u)_{i,j}^n = \text{the right-hand side of (13),} \\ \text{and} \\ (M_2 u)_{i,j}^n = A_2 u_{i,j}^n, \\ (M_3 u)_{i,j}^n = A_3 u_{i,j}^n. \end{array} \right. \quad (17)$$

where A_2 and A_3 are given as in (12). In addition, we define the operator Q on E by components:

$$(Qu)_{i,j}^n = \min_{1 \leq k \leq 3} \{ (M_k u)_{i,j}^n \} \quad (18)$$

By the above definitions, in contrast with the system (12), we get the following discrete approximation for the problem (5):

$$\left\{ \begin{array}{l} \text{(a) } u \leq M_k u, \quad k=1,2,3, \\ \text{(b) } Qu = u, \\ \text{(c) } u|_{\partial N} = f, \\ \text{(d) } u = g \quad \text{for } i=\pm M \text{ or } j=\pm M, \end{array} \right. \quad (19)$$

where, by convention, the first two relations hold component-wise, $f = \{f_{i,j}\}$ and $g = \{g_{i,j}\}$.

As to be explained later, the difference-inequality (13) yields a maximum principle. Consequently there exists a unique solution to the problem (19) which can be constructed by successive approximations. This forms the basis for our second numerical algorithm, an iterative scheme. Given $u^{(0)}$ in E , we denote by $u^{(k)} = \{u_{i,j}^{n,k}\}$ its k -th iterate. The iteration procedure runs as follows:

1.) To start the process, assume the values of $u^{(0)}$ are given so that they satisfy (19.a,c and d).

2.) As the first iterate, set $u^{(1)} = \{u_{i,j}^{n,1}\}$ with

$$u_{i,j}^{n,1} = \begin{cases} \{f_{ij}\} & \text{for } n=N; i,j=0,\pm 1,\dots,\pm M, \\ \{g_{i,j}\} & \text{for } n=0,1,\dots,N-1; i=\pm M \text{ or } j=\pm M, \\ \{Qu^{(0)}\}_{i,j}^n & \text{otherwise.} \end{cases} \quad (20)$$

3.) Suppose we have computed $u^{(k)}$. Then, similar to the above computation, we get

$$u_{i,j}^{n,k+1} = \begin{cases} \{f_{i,j}\} & \text{for } n=N; i,j=0,\pm 1,\dots,\pm M, \\ \{g_{i,j}\} & \text{for } n=0,1,\dots,N-1; i=\pm M \text{ or } j=\pm M, \\ \{Qu\}_{i,j}^n & \text{otherwise.} \end{cases} \quad (21)$$

4.) Preassigned an acceptable error ε . The iteration process terminates at K-th step when $\|u^{(K)} - u^{(K-1)}\| \leq \varepsilon$.

By the properties of the solution of (19), we can show that the iteration-sequence $\{u^{(k)}\}$ defined above converges monotonically in E to the solution u from below. This is a consequence of analytical properties of u to be stated in what follows.

IV. SOME ANALYNICAL RESULTS. We will summarize a few relevant results which enable us to prove the existence, uniqueness of a solution to the problem (19), as well as the convergence of the iteration method proposed in the previous section.

First we state the announced maximum principle:

(R1.) Let $u \in E$ satisfy $u \leq Q(u)$ in $\overset{\circ}{Q}_{M,N}$, the interior of $Q_{M,N}$. Then u cannot attain its maximum in $\overset{\circ}{Q}_{M,N}$.

As an immediate consequence, we have

(R2.) Assume the existence of a solution to (19). Then the solution must be unique. If the data f and g are bounded and positive, then so does the solution u . In fact we get $0 \leq u \leq \max \{ \|f\|, \|g\| \}$.

To prove the existence of a solution, we need

(R3.) Suppose $u \leq Q(u)$ for some $u \in E$. Then the vector $v = Q(u)$ satisfies $v \leq Q(v)$.

The above result implies that the sequence $u^{(k)}$ of iterates defined in (20) and (21) converges, that is

(R4.) Let $\{u^{(k)}\}$ be defined as in (20) and (21). Then the sequence is monotonically increasing and it will converge in E to the solution u of (19).

As a corollary of (R4.), we conclude that

(R5.) The iterative numerical method for the discrete problem (19), as described in §IV, is convergent and stable.

V. NUMERICAL EXAMPLE. As an example, we have carried out extensive numerical computations for the special case when the damping coefficient $p=0$, the spring constant $q=0.2$, the noise level $r=0.2$ and the unit fuel cost $c=0.2$. Also we chose $\Delta t=0.04$, $N=4$ and $\Delta x=\Delta y=0.2$ with $M=6$. The terminal value $f(x,y)=x^2+y^2$, while the tolerable error $\varepsilon=0.5 \times 10^{-4}$. Under these conditions, it is found that the iterative method described in §III converges rather rapidly. Within the specified error ε , most results can be obtained in less than 30 iterations. In Table 1, we show a typical set of computational results for the optimal cost function $u_{i,j}^n$. The results are grouped in three blocks corresponding to $n=1,2,3$. In each block, the entries in the horizontal directions give the values of $u_{i,j}^n$ for i ranging from -6 to $+6$, and the vertical entries are its values for $j=-6, -5, \dots, 5, 6$. Also the regions

corresponding to the continuation sets, in which u satisfies $M_1 u = u$, and their complementary regions are displayed in Figures 2-4, corresponding to $n=1,2,3$. In the figure, the continuation set is marked by the plus (+) signs, while its complement by the star (*) signs. The curves between them are the free boundaries. These figures are important in that they provide a control chart to implement the optimal policy. According to our previous result [1], the optimal policy is to do nothing if the initial state lies in the continuation set, and to make a vertical jump to the boundary otherwise. Subsequently one applies a control only when the state reaches a free boundary by a vertical reflection. In view of the figures, we observe the drastic change in the shape of free boundaries, say for $n=1$ and 3 . For $n=1$, the initial correction is almost one-sided for saving the cost. But near the end, $n=3$, the best policy is a two-sided correction to bring the final state to the origin as close as possible.

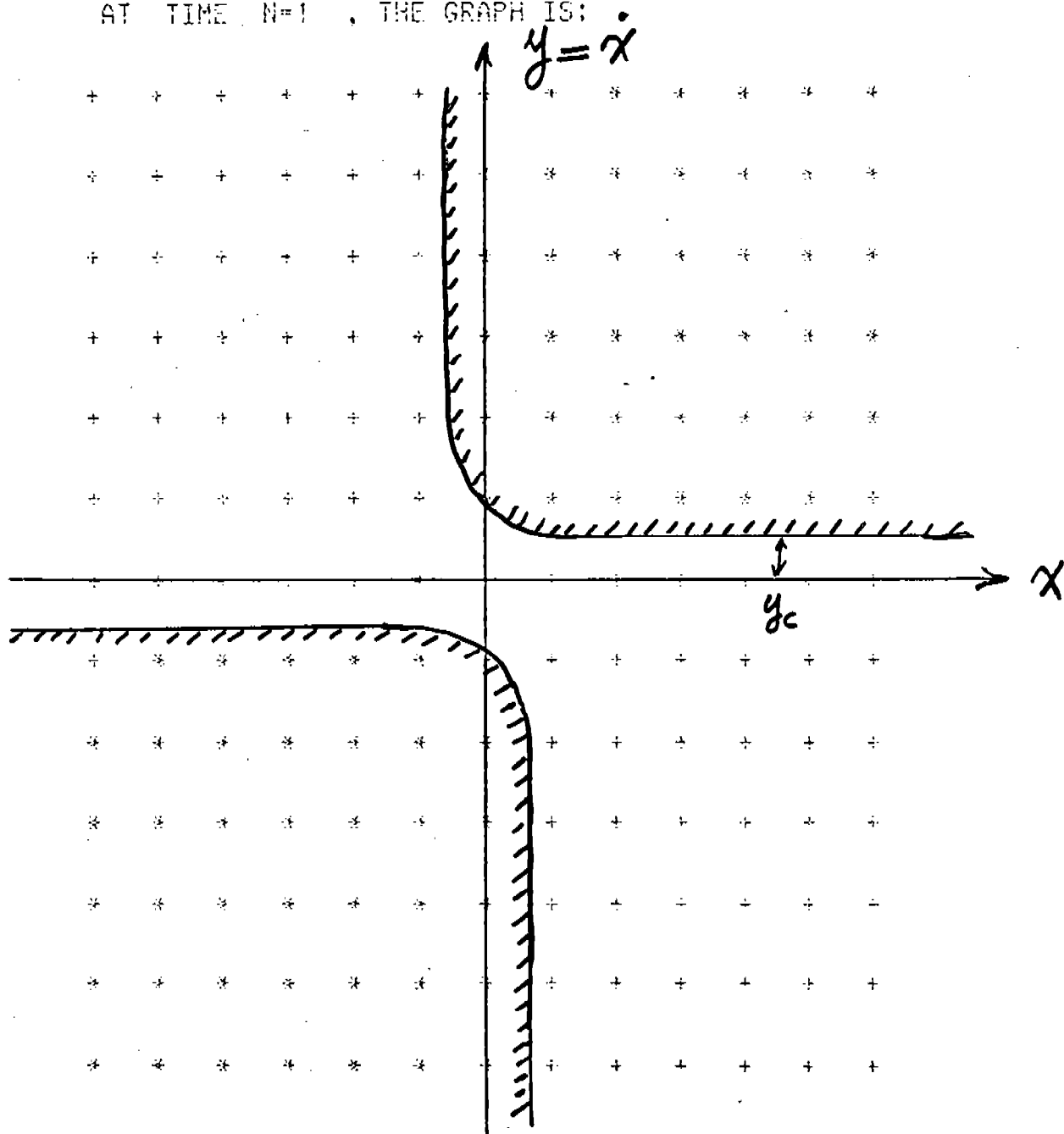
NUMBER OF ITERATION TIMES: 24 356 ARE OBTAINED													
NEW VALUE OF U													
x →													
n=1	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
	-0.0000	1.0000	0.8294	0.5525	0.3556	0.2386	0.2016	0.2416	0.3616	0.5616	0.8415	1.0000	-0.0000
	-0.0000	1.0000	0.7894	0.5125	0.3156	0.1986	0.1616	0.2016	0.3216	0.5216	0.8015	1.0000	-0.0000
	-0.0000	1.0000	0.7494	0.4725	0.2756	0.1586	0.1216	0.1616	0.2816	0.4816	0.7615	1.0000	-0.0000
	-0.0000	1.0000	0.7094	0.4325	0.2356	0.1186	0.0816	0.1216	0.2416	0.4416	0.7215	1.0000	-0.0000
	-0.0000	1.0000	0.6694	0.3925	0.1956	0.0786	0.0416	0.0816	0.2016	0.4016	0.6815	1.0000	-0.0000
	-0.0000	0.9982	0.6415	0.3616	0.1616	0.0416	0.0016	0.0416	0.1616	0.3616	0.6415	0.9982	-0.0000
	-0.0000	1.0000	0.6815	0.4016	0.2016	0.0816	0.0416	0.0786	0.1956	0.3925	0.6694	1.0000	-0.0000
	-0.0000	1.0000	0.7215	0.4416	0.2416	0.1216	0.0816	0.1186	0.2356	0.4325	0.7094	1.0000	-0.0000
	-0.0000	1.0000	0.7615	0.4816	0.2816	0.1616	0.1216	0.1586	0.2756	0.4725	0.7494	1.0000	-0.0000
n=2	-0.0000	1.0000	0.8015	0.5216	0.3216	0.2016	0.1616	0.1986	0.3156	0.5125	0.7894	1.0000	-0.0000
	-0.0000	1.0000	0.8415	0.5616	0.3616	0.2416	0.2016	0.2386	0.3556	0.5525	0.8294	1.0000	-0.0000
	0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
	-0.0000	0.9625	0.7777	0.5166	0.3353	0.2328	0.2032	0.2431	0.3630	0.5629	0.8427	0.9904	-0.0000
	-0.0000	0.9638	0.7474	0.4832	0.2989	0.1938	0.1632	0.2031	0.3230	0.5229	0.8028	0.9897	-0.0000
	-0.0000	0.9682	0.7175	0.4502	0.2627	0.1549	0.1232	0.1631	0.2830	0.4829	0.7628	0.9889	-0.0000
	-0.0000	0.9725	0.6880	0.4174	0.2268	0.1161	0.0832	0.1231	0.2430	0.4429	0.7228	0.9870	-0.0000
	-0.0000	0.9768	0.6589	0.3851	0.1913	0.0774	0.0432	0.0831	0.2030	0.4029	0.6828	0.9841	-0.0000
	-0.0000	0.9811	0.6428	0.3629	0.1630	0.0431	0.0032	0.0431	0.1630	0.3629	0.6428	0.9811	-0.0000
n=3	-0.0000	0.9841	0.6828	0.4029	0.2030	0.0831	0.0432	0.0774	0.1913	0.3851	0.6589	0.9768	-0.0000
	-0.0000	0.9870	0.7228	0.4429	0.2430	0.1231	0.0832	0.1161	0.2268	0.4174	0.6880	0.9725	-0.0000
	-0.0000	0.9889	0.7628	0.4829	0.2830	0.1631	0.1232	0.1549	0.2627	0.4502	0.7175	0.9682	-0.0000
	-0.0000	0.9897	0.8028	0.5229	0.3230	0.2031	0.1632	0.1938	0.2989	0.4832	0.7474	0.9638	-0.0000
	-0.0000	0.9904	0.8427	0.5629	0.3630	0.2431	0.2032	0.2328	0.3353	0.5166	0.7777	0.9625	-0.0000
	0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
	-0.0000	0.9235	0.7292	0.4838	0.3176	0.2282	0.2048	0.2446	0.3644	0.5641	0.8436	0.9796	-0.0000
	0.0000	0.9273	0.7073	0.4558	0.2839	0.1899	0.1648	0.2046	0.3244	0.5241	0.8037	0.9776	-0.0000
	-0.0000	0.9366	0.6867	0.4290	0.2510	0.1517	0.1248	0.1646	0.2844	0.4841	0.7637	0.9756	-0.0000
n=3	-0.0000	0.9448	0.6670	0.4028	0.2186	0.1138	0.0848	0.1246	0.2444	0.4441	0.7237	0.9734	-0.0000
	-0.0000	0.9530	0.6485	0.3778	0.1871	0.0763	0.0448	0.0846	0.2044	0.4041	0.6837	0.9679	-0.0000
	-0.0000	0.9610	0.6437	0.3641	0.1644	0.0446	0.0048	0.0446	0.1644	0.3641	0.6437	0.9610	-0.0000
	-0.0000	0.9679	0.6837	0.4041	0.2044	0.0846	0.0448	0.0763	0.1871	0.3778	0.6485	0.9530	-0.0000
	-0.0000	0.9734	0.7237	0.4441	0.2444	0.1246	0.0848	0.1138	0.2186	0.4028	0.6670	0.9448	-0.0000
	-0.0000	0.9756	0.7637	0.4841	0.2844	0.1646	0.1248	0.1517	0.2510	0.4290	0.6867	0.9366	-0.0000
	-0.0000	0.9776	0.8037	0.5241	0.3244	0.2046	0.1648	0.1899	0.2839	0.4558	0.7073	0.9273	-0.0000
	-0.0000	0.9796	0.8436	0.5641	0.3644	0.2446	0.2048	0.2282	0.3176	0.4838	0.7292	0.9235	-0.0000
	0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000

↑
y

Table 1

Computed values of $u_{i,j}^n$ after 24 iterations

AT TIME $N=1$, THE GRAPH IS:



COSTANTS: $C = 0.2$, $R = 0.2$, $Q = 0.2$

$\Delta X = \Delta Y = 0.2$

$\Delta T = 0.04$

Figure 2

Free Boundary at $n=1$

AT TIME $N=2$, THE GRAPH IS:

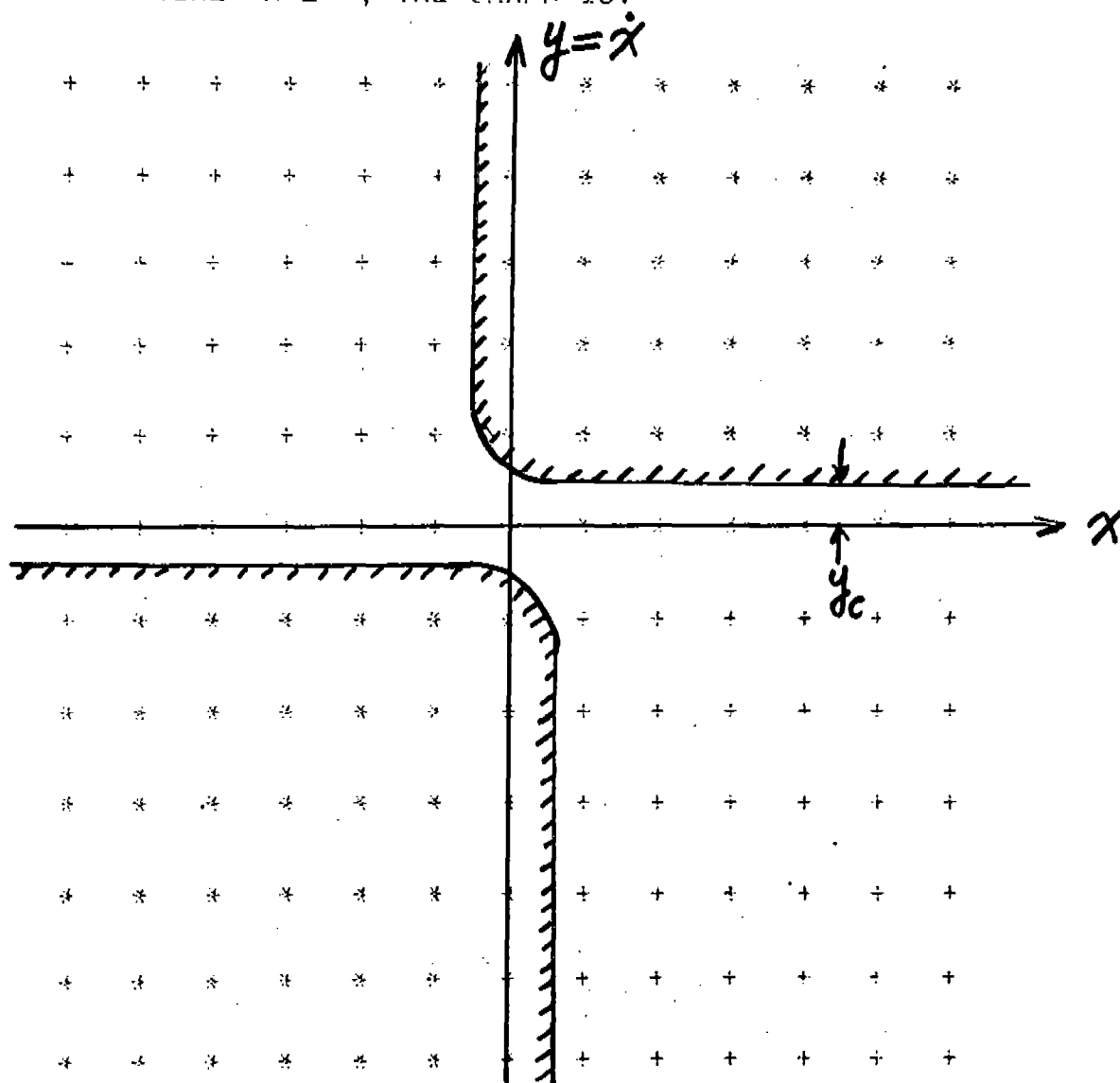


Figure 3

Free Boundary at $n=2$

AT TIME $N=3$ -, THE GRAPH IS:

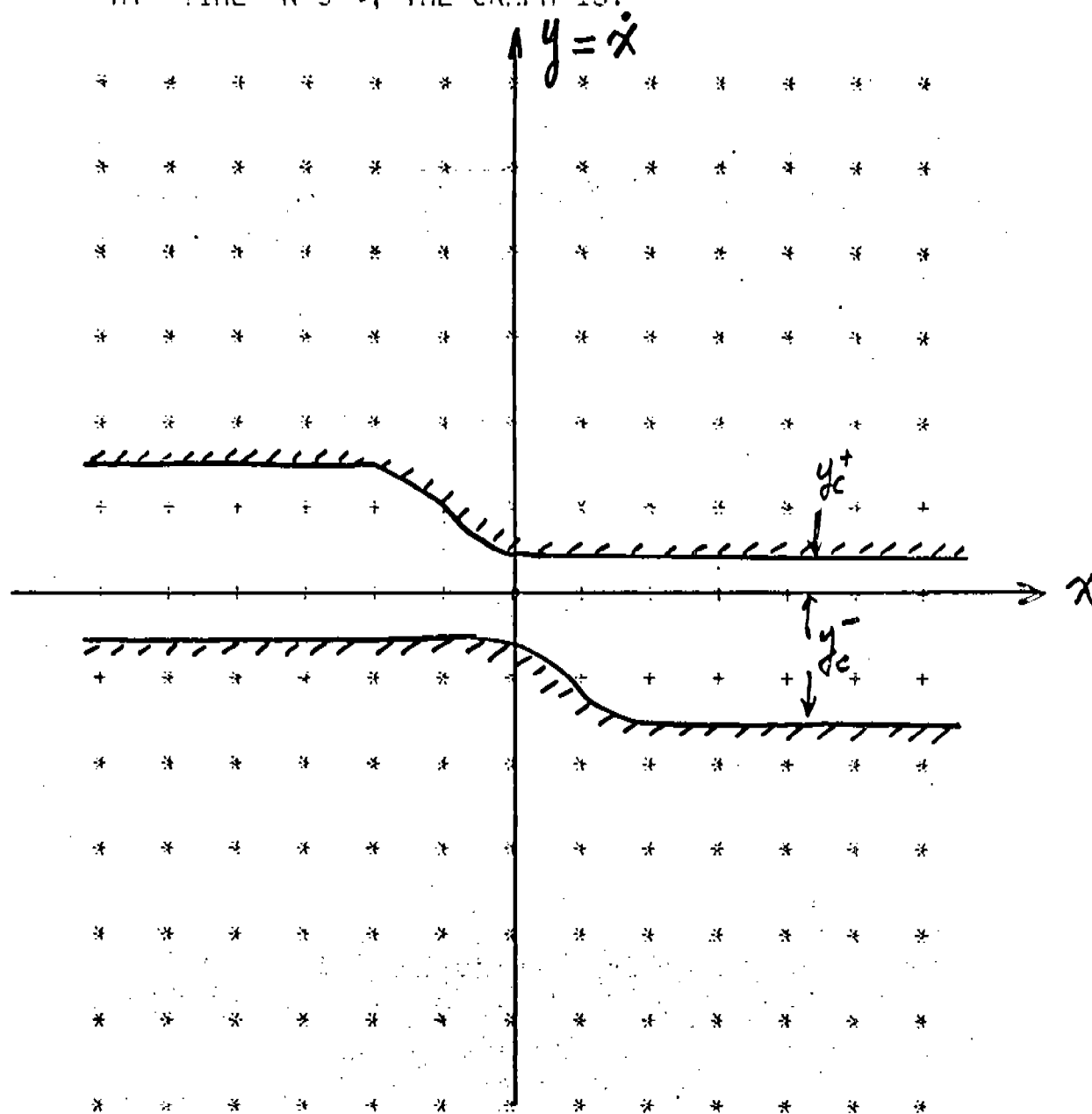


Figure 4

Free Boundary at $n=3$

References

- [1] P.L. Chow and J.L. Menaldi, Optimal Corrections of a Damped Linear Oscillator under Random Perturbations. Trans. of the Second Army Conf. on Appl. Math. and Computing, (1984), pp. 149-158.
- [2] P.L. Chow and J.L. Menaldi, On the Control of a Linear Stochastic System with Finite Horizon, Trans. of the First Army Conf. on Appl. Math. and Computing, (1983), pp. 301-310.
- [3] P.L. Chow and J.L. Menaldi, Additive Control of Stochastic Linear Systems with Finite Horizon, SIAM J. Control and Optim., to appear.
- [4] R. Gonzolez and E. Rofman, On Deterministic Control Problems: An Application Procedure for the Optimal Cost, Research Rept. No. 151, INRIA, Le Chesnay, France, 1982.

ERROR ESTIMATE FOR THE NUMERICAL SOLUTION OF A STOCHASTIC
CONTROL PROBLEM*

P.L. Chow and J.L. Menaldi

Department of Mathematics, Wayne State University

Detroit, Michigan 48202

ABSTRACT. The numerical solution of an optimal stopping problem for a diffusion model is treated. To discretize the problem we use a finite-difference method in such a way that the probabilistic structure is preserved. That is, we regard the discrete system probabilistically, as a Markov chain to approximate the original diffusion process. Thereby the strong convergence and the error of approximation are obtained for the numerical solution of the optimal cost function.

1. INTRODUCTION. We are interested in solving numerically the Hamilton-Jacobi-Bellman equation arising from stochastic control problems. As a first step we will treat a relatively simpler problem of optimal stopping for a diffusion model. For such stochastic control problem, one is referred to the books by Bensoussan and Lions [1] and by Shiriyayev [2], among others, and to the paper by Menaldi [3] for the case of degenerate diffusion.

The main objective of the paper is to study the numerical solution of the optimal cost function, which satisfies a variational inequality as to be shown. In particular we wish to get an error estimate for the approximate solution by using a finite-difference scheme. This scheme is devised in such a way that the probabilistic structure of the problem is preserved. It will be compatible with the approximating Markov chain model to the original diffusion process.

*This work was supported by the ARO Contract DAAG29-83-K-0014.

The idea of using a Markov chain approximation is not new. Among others, Kushner [4] has used this approach extensively in the context of stochastic control. However, he adopted the method of weak convergence or the martingale formulation in the approximation theory. In contrast, we will be concerned with an error estimate for some approximate solution in a strong sense. As far as we know, this kind of result has not been obtained before.

II. DIFFUSION MODELS. The state of the dynamical system under consideration is governed by a stochastic differential equation of Itô type. It reads

$$(1) \quad \begin{cases} dy(t) = \sigma_0[y(t)]dt + \sum_{k=1}^q \sigma_k[y(t)]dW_k(t), & t \geq 0, \\ y(0) = x \end{cases}$$

which is defined in a q -dimensional Wiener space $(\Omega, \mathcal{F}, \mathcal{F}^t, P, W(t), t \geq 0)$, i.e., (Ω, \mathcal{F}, P) is a completed probability space and $W(t) = (W_k(t), k=1, \dots, q)$ is a q -dimensional standard Wiener process with respect to $\{\mathcal{F}^t, t \geq 0\}$. The coefficients $\sigma_k = (\sigma_{ik}(x), i=1, \dots, d)$, $k=1, \dots, q$, are given Lipschitz continuous functions on \mathbb{R}^d , i.e.,

$$(2) \quad \sum_{k=0}^q |\sigma_k(x) - \sigma_k(x')| \leq C|x-x'|, \text{ for } x, x' \text{ in } \mathbb{R}^d,$$

where $|\cdot|$ denotes the Euclidian norm in \mathbb{R}^d . In the equation (1), x is the initial state at the time $t=0$.

Suppose that the only control we have on the system is to decide whether or not we should stop the evolution of the state. This decision should be adapted to the observation of the state, which is assumed to be the actual state of the system and is completely observable. Therefore, if $\tau = \tau(\omega)$, $\omega \in \Omega$, is the random time at which we decide to stop the system, it is a stopping time with respect to $\{\mathcal{F}^t, t \geq 0\}$, i.e.,

$$(3) \quad \{\tau \leq t\} \in \mathcal{F}^t, \quad \text{for every } t \geq 0.$$

Note that $\tau \geq 0$ and $\tau(\omega)$ may be infinite for some $\omega \in \Omega$.

Associated with each decision, or each stopping time τ , we introduce an average cost functional

$$(4) \quad J(\tau) = E \left\{ \int_0^{\tau \wedge T} f[y(t)] dt + h[y(\tau)] \mathbb{1}(\tau < T) \right\},$$

where f and h are given functions. They represent the unit operating cost and the cost of stopping the system, respectively. The horizon T may be finite or infinite. Sometime it can even be random, e.g.,

$$(5) \quad T = \inf \{ t \geq 0 : y(t) \notin \bar{\Theta} \}$$

which is the first exit time of the process $y(t)$ from some closed subset $\bar{\Theta}$ of \mathbb{R}^d . The function $\mathbb{1}(\tau < T)$ equals to 1 if $\tau < T$ and 0 otherwise.

The control problem is to choose τ so that the average cost $J(\tau)$ is minimal. This optimal cost $\hat{u}(x)$ is defined by

$$(6) \quad \hat{u}(x) = \inf \{ J_x(\tau) : \tau \text{ satisfying (3)} \}.$$

where J is written as J_x to show its dependence on x .

Assuming then \hat{u} is finite and smooth, we apply the method of dynamic programming to obtain the equation for optimality

$$(7) \quad \max \{ A\hat{u} - f, \hat{u} - h \} = 0 \quad \text{in } \Theta,$$

while

$$(8) \quad A\hat{u}(x) = -\frac{1}{2} \sum_{i,j=1}^d \left\{ \sum_{k=1}^n \sigma_{ik}(x) \sigma_{jk}(x) \right\} \partial_{ij} \hat{u}(x) - \sum_{i=1}^d \sigma_{i0}(x) \partial_i \hat{u}(x),$$

with $\partial_i = \partial/\partial x_i$, $\partial_{ij} = \partial^2/\partial x_i \partial x_j$. The equation (7) is commonly referred to as a variational inequality. In case that T is given by (5), we must add to the equation (7) the boundary condition

$$(9) \quad \hat{u} = 0 \quad \text{on} \quad \partial\mathcal{O}.$$

Suppose that \hat{u} be the solution of the equation (7) subject to the boundary condition (9). Then we define the continuation set $[\hat{u} < h] = \{x \in \mathcal{O} : \hat{u}(x) < h(x)\}$, which determines the optimal stopping time $\hat{\tau}$. Since, in view of (7), $\hat{u} < h$ implies $A\hat{u} = F$, we have

$$[\hat{u} < h] \subset [A\hat{u} = f],$$

which means that, in the continuation set, it is cheaper to let the system evolve freely without stopping. But, as soon as the state leaves the continuation set, we must stop the system immediately to avoid a higher cost.

In what follows, a finite-difference scheme for solving the variational inequality (7) in \mathcal{O} subject to the condition (9) will be proposed. Probabilistically the diffusion process will be replaced by an appropriate Markov chain so that the structure of the problem is preserved. As mentioned in the previous section, our real interest goes beyond the present problem. Hopefully, this approach may be extended to treat a certain class of nonlinear problems for which the coefficients f and σ could depend on a control parameter, say α , so that $A = A(\alpha)$ and the equation (7) becomes

$$(10) \quad \max_{\alpha} \{ \max [A(\alpha)\hat{u} - f(\alpha)], \hat{u} - h \} = 0.$$

This problem will be studied in the future.

III. MARKOV-CHAIN MODELS. Let Δt denote a small unit of time. The state of the dynamical system at the discrete times $n\Delta t$, $n=0,1,\dots$, is given by $Z(n, \Delta t)$, which evolves according to the equation

$$(11) \quad \begin{cases} Z(n+1) &= Z(n) + \sum_{k=0}^q \sigma_k [Z(n)] \xi_k^{n+1}, \\ Z(0) &= x, \end{cases}$$

where $\xi^n = (\xi_0^n, \xi_1^n, \dots, \xi_q^n)$ with its components being independent, identically distributed random variables, for $n=1, 2, \dots$, and $\sigma_k = (\sigma_{ik}(x), i=1, 2, \dots, d)$, $k=1, 2, \dots, q$, are given as before.

Here a stopping time, the control variable, is an integer-valued random variable $v = v(\omega)$ satisfying

$$(12) \quad \{v \leq n\} \in F^n, \quad n=1, 2, \dots,$$

where F^n is the σ -algebra generated by the variables $\{\xi^1, \xi^2, \dots, \xi^n\}$.

The average cost associates with each control v is given by

$$(13) \quad J_x(v, \Delta t) = E \left\{ \sum_{n=0}^{(v \wedge N)-1} f[Z(n)\Delta t + h[Z(n)]1(v < N)] \right\},$$

where the horizon N is a positive integer defined by

$$(14) \quad N = \inf \{n \geq 0 : Z(n) \notin \bar{\Theta}\}.$$

Again an application of Dynamic Programming yields

$$(15) \quad \max \{A_{\Delta t} \hat{u}_{\Delta t} - f, \hat{u}_{\Delta t} - h\} = 0 \quad \text{in } \Theta.$$

Here we set

$$(16) \quad A_{\Delta t} u(x) = \frac{1}{\Delta t} \left\{ u(x) - E[u(x) + \sum_{k=0}^q \sigma_k(x) \xi_k] \right\},$$

and $\hat{u}_{\Delta t}(x)$ is the optimal cost, i.e.,

$$(17) \quad \hat{u}_{\Delta t}(x) = \inf \{J_x(v, \Delta t) : v \text{ satisfying (12)}\}.$$

The boundary condition for the Equation (15) is

$$(18) \quad \hat{u}_{\Delta t} = 0 \quad \text{on } \partial\Theta.$$

Let α_0, α_1 and β_0, β_1 be positive numbers such that

$$(19) \quad 2q\alpha_1 + \alpha_0 = \alpha_0\beta_0 = \alpha_1\beta_1 = 1.$$

Also we choose the probability distribution of $\xi^n = (\xi_0^n, \xi_1^n, \dots, \xi_q^n)$ to satisfy the following

$$(20) \begin{cases} \xi_k^n \text{ and } \xi_\ell^n \text{ have disjoint supports for } k \neq \ell, \\ P\{\xi_0^n = \beta_0 \Delta t\} = \alpha_0, \\ P\{\xi_k^n = \pm \sqrt{\beta_1 \Delta t}\} = \alpha_1, \text{ for } k=1, \dots, q. \end{cases}$$

Then the operator $A_{\Delta t}$ can be expressed as

$$(21) \quad A_{\Delta t} u(x) = -\frac{\alpha_1}{\Delta t} \sum_{k=1}^q [u(x + \sqrt{\beta_1 \Delta t} \sigma_k(x)) - 2u(x) + u(x - \sqrt{\beta_1 \Delta t} \sigma_k(x))] - \frac{\alpha_0}{\Delta t} [u(x + \beta_0 \Delta t \sigma_0(x)) - u(x)]$$

Note that for a smooth function $u(x)$ we have

$$(22) \quad A_{\Delta t} u(x) = -\frac{1}{2} \sum_{i,j=1}^d \sum_{k=1}^q \int_0^1 \int_{-t}^t \sigma_{ik}(x) \sigma_{jk}(x) \partial_{ij} u(x + s \sqrt{\beta_1 \Delta t} \sigma_k(x)) ds - \sum_{j=1}^d \int_0^1 \sigma_{j0}(x) \partial_j u(x + t \beta_0 \Delta t \sigma_0(x)) dt.$$

One of the key properties of the operator $A_{\Delta t}$ is the validity of the maximum principle which says

$$(23) \begin{cases} \text{if } u(x) \text{ attains a local maximum at } x_0, \text{ then} \\ A_{\Delta t} u(x_0) \leq 0 \text{ for a sufficiently small } \Delta t. \end{cases}$$

Now, in view of (22), we deduce that, if the second derivatives of u are continuous in a neighborhood of x ,

$$(24) \quad |A_{\Delta t} u(x) - Au(x)| \rightarrow 0 \text{ as } \Delta t \rightarrow 0.$$

In passing we remark that the alternative form (21) for $A_{\Delta t}$ has a advantage over the conventional finite-differencing. This enables us to reduce

the number of coupling among the equations from an order of 2^d to that of $2d$. From a numerical point of view, this reduction is significant.

IV. APPROXIMATION RESULTS. Having described two stochastic control models in the previous two sections, we now come to examine the approximation problem. That is, if we approximate the diffusion model by a Markov chain model, what is the approximation error? In particular we are interested in the error estimate for computing the optimal cost by a finite-difference scheme. Our approach to this problem has two distinct features. On one hand, we exploit the analytical characterization of the optimal cost, i.e., the variational inequality or the Hamilton-Jacobi-Bellman equation by replacing the differential operator A by a finite-difference operator $A_{\Delta t}$. On the other hand, we look at the optimal cost itself through an appropriate approximation of the state equation. Specifically, the approximation involves replacing the Brownian motion by a suitable random walk.

There are many published papers concerning the above-mentioned approximation problem. But, to our knowledge, none has addressed to the strong (versus weak) convergence in the approximation. Since we wish to eventually include the deterministic control problems in our analysis, it will be necessary to deal with controlled diffusion processes with possible degeneracy. This is one of the reasons why we are interested in seeking an approximation that will yield the following kind of error estimate

$$(25) \quad E \left\{ \sup_{0 \leq t \leq T} |y(t) - y^{\Delta t}(t)| \right\} \leq C_T (\Delta t)^{\frac{1}{2}},$$

where $T > 0$ is finite, C_T a positive constant depending on σ and T , and $\{y^{\Delta t}(t), t \geq 0\}$ is an approximation process of $\{y(t), t \geq 0\}$ constructed from the Markov chain (11).

We remark that once an estimate like (25) has been established, a corresponding estimate for the optimal cost functions (6) and (17) follows immediately. Therefore, we shall only be concerned with an estimate of the type (25) by giving some probabilistic arguments.

The construction of the process $\{y^{\Delta t}(t), t \geq 0\}$ is suggested by Skorokhod [5], known as the Skorokhod representation. On the same Wiener space where the stochastic equation (1) is based, we define, by induction, the random variables $\eta^n = (\eta_0^n, \eta_1^n, \dots, \eta_q^n)$ and $\vartheta^n = (\vartheta_0^n, \vartheta_1^n, \dots, \vartheta_q^n)$ as follows

$$(26) \quad \begin{cases} \tau_k^0 = 0, \quad W_k^0(t) = W_k(t) \quad \text{for } k=1, \dots, q, \\ \tau_k^{n+1} = \inf \{t \geq 0 : |W_k^n(t)| = |\xi_k^{n+1}|\}, \quad n=0, 1, \dots, \\ W_k^{n+1}(t) = W_k^n(t + \tau_k^{n+1}) - W_k^n(\tau_k^{n+1}), \quad n=0, 1, \dots, \\ \vartheta_k^n = \tau_k^1 + \dots + \tau_k^n, \quad n=1, 2, \dots, \\ \eta_k^{n+1} = W_k^n(\tau_k^{n+1}) = W_k(\vartheta_k^{n+1}) - W_k(\vartheta_k^n), \quad n=0, 1, \dots, \end{cases}$$

and

$$(27) \quad \begin{cases} \tau_0^n = \eta_0^n = \xi_0^n, \\ \vartheta_0^n = \tau_0^1 + \dots + \tau_0^n, \quad n=1, 2, \dots \end{cases}$$

It is possible to show that the random variables η^n have the same probability distribution (20) for the random variables ξ^n , $n=1, 2, \dots$, and that

$$(28) \quad E \left\{ \sum_{k=0}^q \vartheta_k^n \right\} = n \Delta t.$$

Consequently we can replace the equation (11) by

$$(29) \quad \begin{cases} Z(n+1) = Z(n) + \sum_{k=0}^q \sigma_k[Z(n)] \eta_k^{n+1}, & n=0,1,\dots, \\ Z(0) = x, \end{cases}$$

which represents the same Markov chain.

Now we define the approximating process $\{y^{\Delta t}(t), t \geq 0\}$ by

$$(30) \quad y^{\Delta t}(t) = Z(n) \quad \text{if} \quad \sum_{k=0}^q \vartheta_k^n(\omega) \leq t < \sum_{k=0}^q \vartheta_k^{n+1}(\omega).$$

By construction this process is adapted and piecewise constant so that the equation (29) can be written as

$$(31) \quad \begin{cases} y^{\Delta t}(t, \omega) = x + \int_0^{\vartheta_0^n} \sigma_0[y^{\Delta t}(s)] ds \\ \quad + \sum_{k=1}^q \int_0^{\vartheta_k^n} \sigma_k[y^{\Delta t}(s)] dW_k(s), \\ \quad \text{if} \quad \sum_{k=0}^q \vartheta_k^n(\omega) \leq t < \sum_{k=0}^q \vartheta_k^{n+1}(\omega). \end{cases}$$

With aid the above representation, we are able to verify the following result.

Theorem: Under the assumption (2) and with the representation (31), for any given numbers $p > 1$ and $T > 0$, there exists a constant $C = C(p, T) > 0$, depending on p and T , the same constant as in (2), such that

$$(32) \quad E \left\{ \sup_{0 \leq t \leq T} |y_x(t) - y_x^{\Delta t}(t)|^p \right\} \leq C(1 + |x|^p) (\Delta t)^{p/2}$$

for any x in \mathbb{R}^d and $0 < \Delta t \leq 1$.

Here the proof of the theorem will be omitted. It can be found in a forthcoming paper by the authors [6].

In terms of practical computation, we note that the variable x in the discretized equation is still continuous. To discretize the variable x , we may proceed as follows. First, select a convenient basis $\{e_k(x), k=1, 2, \dots, \}$ in a suitable function space, and write

$$(33) \quad u(x) \sim \sum_{k=1}^{\infty} \lambda_k(u) e_k(x),$$

where we may take a finite sum from the series. Then by means the equation (15), we obtain a system of complementary inequalities in $(\lambda_1, \dots, \lambda_k)$ in the space \mathbb{R}^k . Getting an error bound by using (33) is much easier than that for the previous case. For instance, if we choose a mesh in \mathbb{R}^d of the size Δt , i.e., $x = i\Delta t$, where $i = (i_1, \dots, i_d)$ with integers as components, then it is possible to slightly modify the Markov chain (29) in such a way that (32) is preserved and the restriction of (29) to the mesh is still a Markov chain. This means that the corresponding variational inequality (15) for the new Markov chain is the restriction of (15) to the mesh. Thereby all variables now become discrete and the estimate (32) holds.

A paper containing the details of the above arguments is under preparation, and the generalization of the present work to other types of control problems is a subject of our current research.

REFERENCES

- [1] A. Bensoussan and J.L. Lions, Applications des Inequations Variationnelles en Controle Stochastic, Dunod, Paris, 1978, (English translation, North-Holland, 1982).
- [2] A.N. Shiriyayev, Optimal Stopping Rules, Springer-Verlag, New York, 1978.
- [3] J.L. Menaldi, On the Optimal Stopping-Time Problem for Degenerate Diffusions, SIAM J. Control and Optim., 18 (1980), pp. 697-721.
- [4] H.J. Kushner, Probability Methods for Approximation in Stochastic Control and for Eliptic Equations, Academic Press, New York, 1977.
- [5] A.V. Skorokhod, Studies in the Theory of Random Processes, Addition-Wesley, Reading, Mass., 1965.
- [6] P.L. Chow and J.L. Menaldi, On the Numerical Solution of an Optimal Stopping-Time Problem, (in preparation).

FINITE DIFFERENCE METHODS FOR ELLIPTIC SYSTEMS

John C. Strikwerda
University of Wisconsin-Madison

ABSTRACT. This paper is an introduction to finite difference methods for elliptic systems of partial differential equations. Elliptic systems arise in many areas of applications such as incompressible fluid flow and elasticity. The theory for elliptic systems is reviewed and the analogous theory for finite difference schemes is presented.

I. INTRODUCTION. Elliptic systems of partial differential equations arise in many areas of science and engineering, and hence it is important to develop numerical methods for their solution. In this paper we discuss finite difference methods for elliptic systems, presenting both theoretical results and results of sample calculations.

We take as our definition of elliptic systems that given by Douglis and Nirenberg (3).

Definition 1.1

A system of partial differential equations

$$(1.1) \quad \sum_{j=1}^n l_{ij}(x, D) u_j(x) = f_i(x), \quad i = 1, \dots, n,$$

where $D = (-i \partial_{x_1}, \dots, -i \partial_{x_d})$, is an elliptic system if there are integers

$(\sigma_i)_{i=1}^n$ and $(\tau_j)_{j=1}^n$ such that

- 1) $\deg l_{ij}(x, \xi) < \sigma_i + \tau_j$
- 2) there are positive constants c and R such that

$$|\det l_{ij}(x, \xi)| > c |\xi|^{2p} \quad \text{for } |\xi| > R \quad \text{and with } 2p = \sum_{i=1}^n \sigma_i + \sum_{j=1}^n \tau_j.$$

The system (1.1) holds in a domain Ω in R^d and the polynomials in ξ , $l_{ij}(x, \xi)$, are continuous on the closure of Ω .

Another way of expressing condition (2) is to let $\hat{l}_{ij}(x, \xi)$ be the sum of terms of $l_{ij}(x, \xi)$ which are homogeneous of degree $\sigma_i + \tau_j$ in ξ . Then $\det \hat{l}_{ij}(x, \xi)$ is a homogeneous polynomial of degree $2p$. Condition (2) is then equivalent to requiring that

$$\det \hat{l}_{ij}(x, \xi) \neq 0 \quad \text{for } \xi \neq 0.$$

Without loss of generality we can assume that $\sigma_i < 0$ and $\tau_j > 1$ for $1 < i, j < n$.

We now present some examples of elliptic systems that occur frequently in applications. Our first example is the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0$$

$$\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = 0$$

with $\sigma_i \equiv 0$ and $\tau_j \equiv 1$. Also, the first-order Poisson equations

$$u + \frac{\partial p}{\partial x} = f_1$$

$$v + \frac{\partial p}{\partial y} = f_2$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

are elliptic with $\sigma_1 = \sigma_2 = -1$, $\sigma_3 = 0$ and $\tau_1 = \tau_2 = 1$, $\tau_3 = 2$ given that $(u, v, p) = (u_1, u_2, u_3)$. The two-dimensional equations of linear elasticity

$$t_{11} - p_1 \frac{\partial u}{\partial x} - p_2 \frac{\partial v}{\partial y} = 0$$

$$t_{12} - p_3 \frac{\partial u}{\partial y} - p_3 \frac{\partial v}{\partial x} = 0$$

$$t_{12} - p_2 \frac{\partial u}{\partial x} - p_1 \frac{\partial v}{\partial y} = 0$$

$$\frac{\partial t_{11}}{\partial x} + \frac{\partial t_{12}}{\partial y} = f_1$$

$$\frac{\partial t_{12}}{\partial x} + \frac{\partial t_{22}}{\partial y} = f_2$$

are an elliptic with $(t_{11}, t_{12}, t_{22}, u, v) = (u_1, u_2, u_3, u_4, u_5)$,

$(\sigma_i)_{i=1}^5 = (-1, -1, -1, 0, 0)$, and $(\tau_i)_{j=1}^5 = (1, 1, 1, 2, 2)$. The constants p_k are given by

$$p_1 = (1 - \eta^2)^{-1}$$

$$p_2 = \frac{1}{2} (1 + \eta)^{-1}$$

$$p_3 = \eta (1 - \eta^2)^{-1}$$

where η is Poisson's ratio.

The final example is the Stokes equations

$$\nabla^2 u - \frac{\partial p}{\partial x} = 0$$

$$\nabla^2 v - \frac{\partial p}{\partial y} = 0$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

where $\sigma_1 = \sigma_2 = 0$, $\sigma_3 = -1$ and $\tau_1 = \tau_2 = 2$ and $\tau_3 = 1$ with $(u, v, p) = (u_1, u_2, u_3)$.

II. REGULARITY ESTIMATES. The prototype of all elliptic systems is the scalar Poisson equation

$$(2.1) \quad \nabla^2 u = f$$

on a domain Ω . The classical Schauder estimates show that the order of differentiation of the solution to (2.1) is two more than the order of differentiation of the data f (Gilbarg and Trudinger (5)). This increase in the differentiability of the solution over the differentiability of the data is a characterizing feature of elliptic systems.

For the case that the elliptic system (1.1) has constant coefficients, we prove an interior regularity estimate that relates the orders of differentiation of the solution to that of the data.

Theorem 2.1

Let

$$(2.2) \quad \sum_{j=1}^n l_{ij}(D)w_j = f_i, \quad i = 1, \dots, n$$

be an elliptic system defined on \mathbb{R}^d with (σ_i) and (τ_j) as in definition 1.1. If $f_i \in H^{r-\sigma_i}(\mathbb{R}^d)$ then $w_j \in H^{r+\tau_j}(\mathbb{R}^d)$ for any real number r , moreover there exists a constant $C(r)$ such that

$$\begin{aligned} \|w\|_{r+\tau} &= \sum_{j=1}^n \|w_j\|_{r+\tau_j} < C(r) \left(\sum_{i=1}^n \|f_i\|_{r-\sigma_i} + \|w\|_0 \right) \\ &= C(r) (\|f\|_{r-\sigma} + \|w\|_0). \end{aligned}$$

Proof

Consider the system (2.2) written in matrix notation

$$L(D)w(x) = f(x).$$

We begin by using the fourier transform to obtain the system

$$L(\xi)w(\xi) = f(\xi).$$

By condition (1) of definition 1.1, for $|\xi| > R$ the matrix $L(\xi)$ can be decomposed as

$$L(\xi) = |\xi|^\sigma \tilde{L}(\xi) |\xi|^\tau$$

where $\tilde{L}(\xi)$ is bounded in norm independently of $|\xi|$ and $|\xi|^\sigma$ and $|\xi|^\tau$ are the diagonal matrices with entries $|\xi|^{\sigma_1}$ and $|\xi|^{\tau_j}$, respectively. Moreover, for $|\xi| > R$, $\tilde{L}(\xi)$ is an invertible matrix with norm bounded independently of $|\xi|$.

Therefore, for $|\xi| > R$ we have

$$|\xi|^\tau \hat{w}(\xi) = \tilde{L}(\xi)^{-1} |\xi|^{-\sigma} \hat{f}(\xi)$$

and for any real number r

$$|\xi|^{\tau+r} \hat{w}(\xi) = \tilde{L}(\xi)^{-1} |\xi|^{r-\sigma} \hat{f}(\xi).$$

By Parseval's relation we then easily obtain

$$\sum_{j=1}^n \|w_j\|_{r+\tau_j} \leq C(r) \left(\sum_{i=1}^n \|f_i\|_{r-\sigma_i} + \|w\|_0 \right)$$

for some constant $C(r)$ depending on $\tilde{L}(\xi)$ but independent of f or w . This completes the proof.

Estimates of the same form as (2.3) hold for elliptic systems with variable coefficients. The theory of pseudo-differential operators can be used to extend the ideas of the above proof to the case with variable coefficients.

From the regularity estimate on \mathbb{R}^d one can obtain interior regularity estimates for more general domains. To do this we consider the system (2.2) on a domain Ω in \mathbb{R}^d . Let $\varphi(x)$ be a C^∞ cut-off function which is unity on a domain Ω_1 and vanishes off a domain Ω_0 with $\Omega_1 \subseteq \bar{\Omega}_0 \subseteq \Omega$. The operator $L(D)$ applied to the function $\tilde{u}(x) = \varphi(x)u(x)$ gives

$$(2.4) \quad L(D)\tilde{u}(x) = \varphi(x)f(x) - M(x,D)u(x)$$

where $M(x,D)$ is a matrix of differential operators such that the (i,j) -th operator has order less than $\sigma_i + \tau_j$. Moreover $M(x,D)$ vanishes outside of Ω_0 . Thus the system (2.4) can be considered as holding on \mathbb{R}^d . The estimate (2.3) then gives

$$\|u\|_{r+\tau, \Omega_1} \leq C(r) (\|f\|_{r-\sigma, \Omega_0} + \|u\|_{r+\tau-1, \Omega_0})$$

where $C(r)$ also depends on φ and its derivatives. Using a sequence of sets such as Ω_1 and Ω_0 we easily obtain the estimate

$$\|u\|_{r+\tau, \Omega_1} \leq C(r) (\|f\|_{r-\sigma, \Omega} + \|u\|_0).$$

This interior regularity estimate shows that, if a solution to the system (2.2) exists on the domain Ω , then the solution u is smooth in the interior of Ω . The degree of smoothness, or differentiability, is dependent on the differentiability of the data.

III. FINITE DIFFERENCE SCHEMES. We now consider finite difference schemes for elliptic systems. We begin by examining the use of central difference formulas to approximate the Cauchy-Riemann equations

$$(3.1) \quad \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0$$

$$(3.2) \quad \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = 0$$

on the unit square.

For boundary data specify $u = x^2 - y^2$ at $x = 0$ and $x = 1$, and specify $v = 2xy$ at $y = 0$ and $y = 1$. Using central difference approximations along with one-sided differences to (3.1) at the boundaries one obtains as the solution

$$(3.3) \quad u_{ij} = x_i^2 - y_j^2 + \epsilon_i$$

$$v_{ij} = 2x_i y_j$$

where $x_i = ih$, $y_j = jh$, and $h = 1/(2M)$. (This example is discussed in more detail in Bube and Strikwerda (2)).

The quantity ϵ_i is given by

$$\epsilon_i = \begin{cases} 0 & i \text{ even} \\ -h^2 & i \text{ odd} \end{cases}$$

Notice that the solution is second-order accurate, but not smooth. In particular, the second divided difference of u_{ij} with respect to x does not converge to the second partial derivative of $u(x)$ with respect to x . This nonconvergence of divided differences does not occur for the usual finite difference schemes for a single second-order elliptic equation. Bramble and Hubbard (1) showed that for solutions of a finite difference scheme approximating a single second order elliptic equation, the divided differences converge to the corresponding partial derivatives, provided the data is sufficiently smooth. Moreover, if the divided difference formulas are accurate enough, the rate of convergence of the divided differences is the same as that of the solution itself. Similar results have been given by Thom  and Westergren (8), Vainikko and Tamme (11), and others.

As the example shows, not all consistent finite difference schemes for elliptic systems have the same regularity property that schemes such as the standard five-point Laplacian scheme have. Bube and Strikwerda (2) showed that a class of finite difference schemes, called regular schemes, do have the regularity property. These regular schemes may be defined as follows.

Definition 3.1.

For $i, j = 1, \dots, n$, let L_{ij} be a difference operator with symbol $\ell_{ij}(h, x, h)$. The system of difference equations

$$(3.4) \quad \sum_{j=1}^n L_{ij} u_j(x) = f_i(x) \quad i = 1, \dots, n,$$

is a regular elliptic system if there are sets of integers $(\sigma_i)_{i=1}^n$ and $(\tau_j)_{j=1}^n$ such that each L_{ij} is a difference operator of order at most $\sigma_i + \tau_j$, and if there are positive constants C, ξ_0, h_0 such that

$$(3.5) \quad |\det \ell_{ij}(h, x, \xi)| > C |\xi|^{2p}$$

for $\xi_0 < \xi < \frac{\pi}{h}$ and $0 < h < h_0$, where $2p = \sum \sigma_i + \sum \tau_j$.

The symbol of a difference operator L is defined by $L e^{ix\xi} = \ell(h, x, h\xi) e^{ix\xi}$, i.e. the symbol is the factor multiplying $e^{ix\xi}$ which results from applying the operator to $e^{ix\xi}$.

One can easily check that the finite difference scheme for the Cauchy-Riemann equations which uses central differences does not satisfy the determinant condition (3.5). The symbol for the central difference scheme for (3.1) is

$$\begin{pmatrix} i \frac{\sin h\xi_1}{h} & -i \frac{\sin h\xi_2}{h} \\ i \frac{\sin h\xi_2}{h} & i \frac{\sin h\xi_1}{h} \end{pmatrix},$$

and the absolute value of the determinant is $(\sin^2 h\xi_1 + \sin^2 h\xi_2)h^{-2}$.

This determinant vanishes for $(\xi_1, \xi_2) = (\pi h^{-1}, 0), (0, \pi h^{-1}),$ and

$(\pi h^{-1}, \pi h^{-1})$. As shown by Bube and Strikwerda (2) the loss of smoothness in the example is a consequence of the vanishing of the determinant.

There are several regular schemes for the Cauchy-Riemann equations. One common regular scheme is the staggered grid method used by Ghil and Balgovind (4), Lomax and Martin (6), and others. A regular scheme using a non-staggered grid is obtained by using the following approximations

$$\frac{\partial u}{\partial x} \sim \delta_{x0} u - \frac{h^2}{6} \delta_{x+}^2 \delta_{x-} u$$

$$\frac{\partial v}{\partial y} \sim \delta_{y0} v - \frac{h^2}{6} \delta_{y+}^2 \delta_{y-} v$$

$$\frac{\partial u}{\partial y} \sim \delta_{y0} u - \frac{h^2}{6} \delta_{y+} \delta_{y-}^2 u$$

$$\frac{\partial v}{\partial x} \sim \delta_{x0} v - \frac{h^2}{6} \delta_{x+} \delta_{x-}^2 v,$$

where the second subscript of 0, +, or - indicates a central, forward, or backward divided difference. If we set

$$\zeta(\xi) = (\sin(h\xi) + \frac{4}{3} e^{ih\xi/2} \sin^3 \frac{1}{2} h\xi) h^{-1}$$

then the symbol of this system of difference equations is

$$\begin{pmatrix} i \zeta(\xi_1) & -i \zeta(\xi_2) \\ i \overline{\zeta(\xi_2)} & i \overline{\zeta(\xi_1)} \end{pmatrix}$$

and the determinant is

$$(|\zeta(\xi_1)|^2 + |\zeta(\xi_2)|^2)$$

which does not vanish for $|\xi_1| + |\xi_2|$ non zero. Research is currently being done on finding an efficient algorithm for solving the system of finite difference equations resulting from this approximation.

A regular finite difference scheme for the Stokes equations has been proposed by Strikwerda (9). This scheme was shown to be second-order accurate on non-orthogonal non-uniform grids. This scheme was also used with the incompressible Navier-Stokes equations, Strikwerda (10) and Nagel and Strikwerda (7).

Currently research is being done to consider the regularity at the boundary of the domain for finite difference schemes. It should be possible to determine the effect of the numerical boundary conditions on the accuracy and smoothness of the solution near the boundary. It is also important to devise better methods for solving the finite difference equations arising from approximations to elliptic systems.

BIBLIOGRAPHY

1. J. H. Bramble and B. E. Hubbard, Approximation of derivatives by finite difference methods in elliptic boundary value problems, *Contributions to Differential Equations*, 3, (1964), pp. 399-410.
2. K. Bube and J. Strikwerda, Interior regularity estimates for elliptic systems of difference equations, *SIAM J. Numer. Anal.*, 20, (1983) pp. 639-656.
3. A. Douglis and L. Nirenberg, Interior estimates for elliptic systems of partial differential equations, *Comm. Pure Appl. Math.*, 8, (1955), pp. 530-538.
4. M. Ghil and R. Balgovind, A fast Cauchy-Riemann solver, *Math. Comp.*, 33, (1979), pp. 585-635.
5. D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlog, New York, 1977.
6. H. Iomax and E. D. Martin, Fast direct numerical solution of the non-homogeneous Cauchy-Riemann equations, *J. Comput. Phys.*, 15, (1974), pp. 55-80.
7. Y. Nagel and J. C. Strikwerda, A numerical study of the flow in a spinning and coning cylinder, in preparation.
8. V. Thomée and B. Westergren, Elliptic difference equations and interior regularity, *Numer. Math.*, 11, (1968), pp. 196-210.
9. J. C. Strikwerda, Finite difference methods for the Stokes and Navier-Stokes equations, *SIMA J. Sci. Statist. Comput.*, 5, (1984), pp. 56-68.
10. J. C. Strikwerda, A numerical study of Taylor-Couette motion, *J. Fluid Mech.*, submitted.
11. G. M. Vainikko and E. E. Tamme, Convergence of a difference method in the problem concerning the periodic solution of an elliptic equation, *USSR Comput. Math. and Math. Phys.* 16, (1976), pp. 105-117.

GENERALIZED ISOVECTORS AND SIMILARITY SOLUTIONS*

Frank B. Estabrook and Hugo D. Wahlquist
Jet Propulsion Laboratory 169-327
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109

ABSTRACT The manipulatory techniques of differential geometry allow systematic derivation of invariance groups of sets of partial differential equations. For each generator or isovector of the algebra of such a group, a family of similarity solutions may be found by integrating a set of equations with one fewer independent variables. The equations for the components of the isovectors themselves are overdetermined linear partial differential equations, which can be solved algorithmically by symbolic manipulation programs. Such families of similarity solutions can be generalized by relaxing the definition of isovector, so that its components satisfy nonlinear equations, and the group property is lost. The resulting "generalized similarity solutions" nonetheless can be useful and physically significant.

I. DIFFERENTIAL GEOMETRY AND APPLIED MATHEMATICS Topics such as partial differential equations (not necessarily linear), Hamiltonian mechanics, variational and perturbational techniques, and Lie group theory can all be treated at a deep level as described by tensor structures on differential manifolds. It should thus be stressed that differential geometry is a much broader discipline than that traditionally taught, which over-emphasized Riemannian geometry. In general, one does not have a metric tensor g_{ij} , one cannot "raise and lower" indices, and one must carefully maintain the distinction between "contravariant" objects such as vector fields, and "covariant" objects such as 1-form fields. Without a metric, one is limited in differentiation operations to the generalized curl, d (applied to completely antisymmetric covariant fields, n -forms), and to the "substantial" or Lie derivative operation along a given vector field. The inner product, or contraction of vectors and forms, and the outer or antisymmetrized tensor product of forms, completes the set of operations: $d, \mathcal{L}_V, \lrcorner, \wedge$. The resulting "exterior calculus" has been expounded in a number of texts, and, in our opinion, will in the near future be routinely taught at the undergraduate level. We recommend a new text by William L. Burke.⁽¹⁾

II. CARTAN - KÄHLER THEORY Sets of first order partial differential equations are treated in Cartan-Kähler theory as defined by ideals of differential forms.⁽²⁾ Such an ideal is generated by a set of forms α^i , or by a set β^i algebraically equivalent to the α^i —i.e., any form, say β , in the second set can be expressed as a sum

over all forms of the same rank in the first ideal, with functions (or scalar fields) as coefficients, and conversely. Any form in the ideal of these generators may be so expressed, and in fact the defining property is that all the forms in the ideal vanish when pulled back (or mapped) onto an "integral" submanifold. The set α^1 should be completed with all other generators having this property of vanishing on any of the integral submanifolds, in particular the forms da^1 . The integral submanifolds are solutions of sets of partial differential equations which appear when a choice of independent and dependent variables is made.

Topics such as the symmetries, potentials, conservation laws, and so on, of a set of p.d.e.'s can all be treated systematically when a closed differential ideal can be found to express it. We believe this approach to be particularly useful when a set is nonlinear (or when it appears to be nonlinear in the variables first given). Consider as an example an 8 dimensional space, with points continuously labeled by coordinates or parameters $U, \Omega, A, B, F, G, \rho, t$, and in it an ideal I generated by a set of two 1-form fields and four 2-form fields:

$$\begin{aligned} dU &= A dt - B dp \\ d\Omega &= F dt - G dp \\ dA \wedge dt + dB \wedge dp \\ dF \wedge dt + dG \wedge dp \\ dB \wedge dt + dA \wedge dp &= [e^{-2U}(F^2 - G^2) - B/\rho] dp \wedge dt \\ dG \wedge dt + dF \wedge dp &= [2(BG - AF) - G/\rho] dp \wedge dt \end{aligned} \quad (1)$$

I is closed, that is $dI \subset I$. Cartan's theory⁽²⁾ allows us to calculate the maximum dimension of integral submanifold to be 2; if we adopt ρ and t as independent variables, the condition that all (the generators) of I vanish reduces to a set of p.d.e.'s that in fact are of great interest in the theory of (nonlinear) cylindrical gravitational waves⁽³⁾:

$$\begin{aligned} U_{\rho\rho} + \frac{1}{\rho} U_{\rho} - U_{tt} &= e^{-2U}(\Omega_t^2 - \Omega_{\rho}^2) \\ \Omega_{\rho\rho} + \frac{1}{\rho} \Omega_{\rho} - \Omega_{tt} &= 2(\Omega_{\rho} U_{\rho} - \Omega_t U_t) \end{aligned} \quad (2)$$

Soliton solutions, and many beautiful properties have recently been discovered for this set. The ideal I is one way of representing this, that allows the underlying group and algebraic structures to be uncovered.

III. CAUCHY CHARACTERISTICS When presented with an ideal I for analysis, the first associated structure to seek is a vector field V that has the property:

$$V \lrcorner I \subset I \quad (3)$$

If it exists, such a field, denoted a Cauchy Characteristic of I , when contracted on any form in I , yields a form (of one less rank) also in I . These equations are an overdetermined homogeneous linear algebraic set for the components of V . An example is the field of flow, or set of trajectories, that belongs to a Hamilton-Jacobi Equation. It is easily shown that any Cauchy Characteristic vector V must lie in the integral manifolds of maximum dimension. A non-trivial further result is then Cartan's theorem, that for any such V a coordinate transformation can be found that eliminates one (independent) coordinate from explicit appearance in the set of generators for I . The solutions thus depend on one less independent variable than at first may appear.

IV. ISOVECTORS The next class of associated vector fields are the isovectors, or generators of Lie's infinitesimal symmetries (the same as treated by Ovsiannikov - Ames, Bluman and Cole, etc.) In our language, they satisfy

$$\mathcal{L}_V I \subset I \quad (4)$$

and can readily be seen to form a group. Now one must integrate a homogeneous, linear, overdetermined set of first order partial differential equations for the components of V . This can be done algorithmically by repeated differentiation (sic!), introducing at each step more functions, of fewer variables. Beautiful and effective symbolic manipulation programs (in the language REDUCE) have recently been developed for this at Twente University, in Holland, c.f. recent theses by P. Gragert and P. Kersten, and also a survey paper with their professor R. Martini.⁽⁴⁾

V. SIMILARITY SOLUTIONS We can now precisely define a similarity solution of an ideal $I = \{\alpha^1\}$, belonging to an isovector V of I (one usually writes the most general V belonging to the entire isogroup), as an integral manifold of an augmented ideal I' generated by forms α^1 and

$$I' = \{\alpha^1, V\lrcorner\alpha^1\} \quad (5)$$

Assuming that I was complete, that is, that $d\alpha^1 \subset I$, it is immediately calculable that I' is also complete, since from the identity $d(V\lrcorner\alpha^1) = \mathcal{L}_V\alpha^1 + V\lrcorner d\alpha^1 \subset I'$. Moreover, I' has an isovector, viz., V —we have been able consistently to impose it. Thus by Cartan's theorem, the number of independent variables in I' can be reduced by one. We will now be searching for a (nonempty!) subset of solutions of I . If the original number of independent

variable was greater than two, and we have more than one independent isovector, this process can be repeated until a set that is readily solvable is obtained, or until a set of ordinary differential equation is at hand. Returning to the above example, one verifies that the following is an isovector:

$$V = t \frac{\partial}{\partial t} + \rho \frac{\partial}{\partial \rho} - A \frac{\partial}{\partial A} - B \frac{\partial}{\partial B} - F \frac{\partial}{\partial F} - G \frac{\partial}{\partial G} \quad (6)$$

(This could easily be found ad hoc by any applied mathematician, as it generates scaling invariance, so we repeat that, in general, one should use a superposition of all generators of the isogroup!) The Cartan coordinate reduction allowed by Eq. (6) can be achieved by introducing new independent variables say $\eta = \rho/t$ (which is such that $f_{\eta}\eta = 0$) and any other independent function of ρ, t . The latter then drops out of I' and one finds a set of o.d.e.'s:

$$\begin{aligned} U''(\eta^2 - 1) + U'(2\eta - \frac{1}{\eta}) &= e^{-2U}(\Omega')^2(1 - \eta^2) \\ \Omega''(\eta^2 - 1) + \Omega'(2\eta - \frac{1}{\eta}) &= 2U'\Omega'(\eta^2 - 1) \end{aligned} \quad (7)$$

These have been solved by quadrature by E. Fischer.(5)

In the same paper, Fischer finds another reduction of Eq. (2) to a set of o.d.e.'s. As is in fact often tried in relativity theory, the ansatz that U and Ω depend only on the combination $\eta = \rho^2 - t^2$ also works! One finds

$$\begin{aligned} 2\eta U'' + 3U' &= -2\eta e^{-2U}(\Omega')^2 \\ 2\eta \Omega'' + 3\Omega' &= 4\eta \Omega' U' \end{aligned} \quad (8)$$

and, again, these are solvable by quadrature. But the resulting solutions do not belong to an infinitesimal symmetry of I as given! What's going on?

VI. GENERALIZED SYMMETRIES First, it must be said that the description of a set of p.d.e.'s by a differential ideal I is not unique. It may even be possible to express I in terms of fewer variables, although there are certain criteria for I to be "well set" that we do not expound here. For any such alternate set, a different but closely related ideal can result, and a systematic understanding of how its isogroup is built into, or onto, that first found, is not clear to us. This is one of the reasons we only claim to be expounding tools for applied mathematicians.

Prolongation of an I by including higher partial derivatives has

been considered carefully by Anderson and Ibragimov⁽⁶⁾, who find, in the limit of infinite prolongation, new symmetries of "Lie-Bäcklund" type related to soliton transformations and other newly appreciated techniques for solution generation.

Prolongation of an I by finding conservation laws, and generalized conservation laws, and introducing new potential-like variables, has been found by us⁽⁷⁾ to be equally effective in finding other classes of generalized symmetries. An example is the Burgers equation, whose linearity (an infinite number of isogroup generators!) is not found until an additional potential variable is included in the set of p.d.e.'s.

Finally, what about the example above, at the end of section V? The ansatz was found by Fischer to result from a "generalized isovector" V satisfying the equation:

$$\mathbb{L}_V I \subset I' \quad (9)$$

This is still sufficient to preserve the closure of I' , using Eq. (5), but such vectors V no longer form a group. If finding similarity-type special solutions is all one wants, this is in principle no problem. What may be a practical problem is that the equations (9), (5), for V are no longer linear in its components. It remains an open question whether such generalized isovectors and their similarity solution families have any significant relation to the other approaches to generalized symmetries we have mentioned above.

*Research sponsored by the U.S. Army Research Office through an agreement with the National Aeronautics and Space Administration.

References

- (1) William L. Burke, Applied differential geometry (Cambridge University Press, 1985).
- (2) F.B. Estabrook, chapter in Backlund Transformations, R.M. Miura, Ed., Lecture Notes in Mathematics No. 515 (Springer-Verlag, 1976); also chapter in Geometrical Approaches to Differential Systems, R. Martini, Ed., Lecture Notes in Mathematics No. 810 (Springer-Verlag, 1980).
- (3) W. Kinnersley, in G. Shaviv and J. Rosen, Eds., Relativity and Gravitation (Wiley, 1975).

- (4) P.K.H. Gragert, P.H.M. Kersten and R. Martini, "Symbolic Computations in Applied Differential Geometry", Acta Applicandae Mathematicae, 1, 43-77 (1983).
- (5) E. Fischer, "Similarity Solutions of the Einstein and Einstein-Maxwell Equations", J. Phys. A: Math. Gen. 13, L81-4 (1980).
- (6) R.L. Anderson and N.H. Ibragimov, Lie-Backlund Transformations and Applications (SIAM, 1982).
- (7) H.D. Wahlquist and F.B. Estabrook, "Prolongation Structures of Nonlinear Evolution Equations, I", J. Math. Phys. 16, 1-7 (1975). F.B. Estabrook and H.D. Wahlquist, "II", J. Math. Phys. 17, 1293-7 (1976).
- (8) B.K. Harrison and F.B. Estabrook, "Geometric Approach to Invariance Groups and Solution of Partial Differential Equations", J. Math. Phys., 12, 653-66 (1971).

APPLICATION OF RECIPROCAL BÄCKLUND TRANSFORMATIONS
TO STEFAN PROBLEMS IN NONLINEAR HEAT CONDUCTION

Colin Rogers*

School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332 USA

ABSTRACT. Reciprocal Bäcklund transformations are used to investigate both one-phase and two-phase Stefan problems in nonlinear heat conduction. A new class of exact solutions to the associated nonlinear moving boundary value problems is derived which is analogous to that obtained by Neumann in linear heat conduction.

I. INTRODUCTION. Storm [1], in an investigation of heat transport in simple metals showed that for an important class of such materials a Bäcklund transformation may be introduced which reduces the prevailing nonlinear heat conduction equation to the classical 1+1 heat equation. The reduction was used to solve a fixed boundary value problem involving the temperature distribution in a half-space with an insulated boundary. It has been shown recently that the Storm transformation may be set in the context of a class of reciprocal Bäcklund transformations which allow the reduction of a wide variety of nonlinear boundary value problems to linear

*Permanent address: Department of Applied Mathematics,
University of Waterloo, Waterloo, Ontario, Canada.

canonical form [2,3]. Here, both one-phase and two-phase Stefan problems are considered for materials of Storm-type. Such moving boundary problems arise naturally in the analysis of melting and solidification processes [4]. Their complexity resides in the fact that the heat balance condition at the moving interface produces a nonlinear boundary condition. Here, there is that additional complication that the heat conduction equations considered are themselves nonlinear. However, with attention restricted to Storm-type materials, it is shown that introduction of a reciprocal transformation allows the construction of a class of exact solutions analogous to the classical Neumann solutions of linear heat conduction.

II. THE RECIPROCAL TRANSFORMATION. In what follows, we make use of the following result [5]:

Theorem

The conservation law

$$\frac{\partial}{\partial t} \{T(\partial/\partial x; \partial/\partial t; u)\} + \frac{\partial}{\partial x} \{F(\partial/\partial x; \partial/\partial t; u)\} = 0 \quad (1)$$

is transformed to the reciprocally associated conservation law

$$\frac{\partial T^*}{\partial t^*} + \frac{\partial F^*}{\partial x^*} = 0 \quad (2)$$

by the reciprocal transformation:

$$\left. \begin{aligned} dx^* &= Tdx - Fdt, & t^* &= t \\ T^* &= \frac{1}{T(D^*; \partial^*; u)}, & F^* &= \frac{-F(D^*; \partial^*; u)}{T(D^*; \partial^*; u)} \end{aligned} \right\} R \quad (3)$$

where

$$D^* := \frac{\partial}{\partial x} = \frac{1}{T^*} \frac{\partial}{\partial x^*}, \quad \partial^* := \frac{\partial}{\partial t} = \frac{F^*}{T^*} \frac{\partial}{\partial x^*} + \frac{\partial}{\partial t^*}$$

$$T(\partial/\partial x; \partial/\partial t; u) := T(u, u_x, u_{xx}, \dots; u_t, u_{tt}, \dots)$$

$$F(\partial/\partial x; \partial/\partial t; u) := F(u, u_x, u_{xx}, \dots; u_t, u_{tt}, \dots)$$

The reciprocal nature of the transformation resides in the involutory property $R^2 = I$.

In particular, the above result shows that the nonlinear equation

$$\frac{\partial}{\partial t} [\Phi(u)] - \frac{\partial}{\partial x} \left[\Phi(u) \sum_{i=1}^n \alpha_i D^i \left(\frac{1}{\Phi(u)} \right) \right] = 0 \quad (4)$$

where $D := \frac{1}{\Phi(u)} \frac{\partial}{\partial x}$ is reducible to the linear canonical form

$$\frac{\partial u^*}{\partial t^*} + \frac{\partial}{\partial x^*} \left[\sum_{i=1}^n \alpha_i \frac{\partial^i u^*}{\partial x^{*i}} \right] = 0 \quad (5)$$

via the reciprocal transformation

$$\left. \begin{aligned} dx^* &= \Phi(u)dx + \Phi(u) \sum_{i=1}^n \alpha_i D^i \left(\frac{1}{\Phi(u)} \right) dt, & t^* &= t \\ u^* &= \frac{1}{\Phi(u)} \end{aligned} \right\} \quad (6)$$

In the sequel, a special case of this result is used in the analysis of a class of Stefan problems in nonlinear heat conduction.

III. A CLASS OF SINGLE PHASE STEPHAN PROBLEMS IN NON-LINEAR HEAT CONDUCTION. The following nonlinear moving boundary value problem is considered:

$$\rho c_p(T) \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} [\kappa(T) \frac{\partial T}{\partial x}], \quad 0 < x < X(t) \quad (7)$$

$$\kappa(T) \frac{\partial T}{\partial x} = U(t) \quad \text{on} \quad x = 0, \quad t > 0 \quad (8)$$

$$\left. \begin{array}{l} \kappa(T) \frac{\partial T}{\partial x} = L\rho\dot{X}(t) \\ T = T_f \end{array} \right\} \quad \text{on} \quad x = X(t) \quad (9)$$

$$X(0) = 0. \quad (10)$$

In the above, $T(x,t)$ denotes the temperature distribution in a medium wherein the specific heat c_p and thermal conductivity κ are temperature dependent. The density ρ of the material is taken to be constant. L denotes the latent heat of fusion of the material and liberation of heat is envisaged to take place during a phase change which occurs at the temperature T_f .

If we set

$$\Phi(T) = \int_{T_0}^T S(\sigma) d\sigma, \quad S = \rho c_p(T) \quad (11)$$

$$(12)$$

then the nonlinear heat equation (7) may be written as

$$\frac{\partial}{\partial t} [\Phi(T)] - \frac{\partial}{\partial x} [\kappa(T) \frac{\partial T}{\partial x}] = 0. \quad (13)$$

The reduction of the previous section with

$$\alpha_i = \begin{cases} 0 & i \neq 1 \\ -\kappa^* & i = 1, \kappa^* > 0 \end{cases} \quad (14)$$

shows that (13) is taken to the 1+1-classical heat equation

$$\frac{\partial T^*}{\partial t^*} = \kappa^* \frac{\partial^2 T^*}{\partial x^{*2}}, \quad (15)$$

via the reciprocal transformation

$$\left. \begin{aligned} dx^* &= \Phi(T) dx + \kappa(T) \frac{\partial T}{\partial x} dt, \quad t^* = t \\ T^* &= \frac{1}{\Phi(T)} \end{aligned} \right\} \quad (16)$$

subject only to the condition

$$\kappa^* \Phi' / \Phi^2 = \kappa(T). \quad (17)$$

The applicability of the condition (17) to simple metals was discussed in [1].

Under the reciprocal transformation (16)

$$\begin{aligned} \frac{\partial x^*}{\partial x} &= \Phi(T), \\ \frac{\partial x^*}{\partial t} &= \kappa(T) \frac{\partial T}{\partial x} = \int_0^x \frac{\partial}{\partial x} [\kappa(T) \frac{\partial T}{\partial x}] dx + \kappa(T) \frac{\partial T}{\partial x} \Big|_{x=0} \\ &= \int_0^x \frac{\partial}{\partial t} [\Phi(T)] dx + U(t) \end{aligned}$$

whence

$$x^*(x, t) = \int_0^x \Phi dx + \Theta(t) - \Theta(0), \quad (18)$$

where $\Theta(t) = U(t)$ and we have taken $x^*(0, 0) = 0$.

Accordingly, the fixed boundary condition (8) becomes

$$\kappa^* \frac{\partial T^*}{\partial x^*} = -UT^* \quad \text{on} \quad x^* = \theta(t^*) - \theta(0). \quad (19)$$

Moreover,

$$\begin{aligned} \frac{\partial x^*}{\partial t} &= \int_{X(t)}^x \frac{\partial}{\partial x} [\kappa(T) \frac{\partial T}{\partial x}] dx + \kappa(T) \frac{\partial T}{\partial x} \Big|_{x=X(t)} \\ &= \int_{X(t)}^x \frac{\partial}{\partial t} [\Phi(T)] dx + L\rho \dot{X}(t) \end{aligned}$$

so that we obtain an alternative expression

$$x^*(x, t) = \int_{X(t)}^x \Phi(T) dx + [\Phi(T_f) + L\rho]X. \quad (20)$$

Thus, the moving boundary conditions (9) become

$$\left. \begin{aligned} \kappa^* \frac{\partial T^*}{\partial x^*} &= -L\rho T^* \dot{X} \\ T^* &= \frac{1}{\Phi(T_f)} \end{aligned} \right\} \text{on } x^* = X^* \quad (21)$$

where

$$X^* = x^* \Big|_{x=X(t)} = [\Phi(T_f) + L\rho]X(t) \quad (22)$$

so that the initial condition (10) becomes

$$X^* \Big|_{t^*=0} = 0. \quad (21)$$

Furthermore, on $x^* = X^*$ it is seen that

$$dx^*/dt^* = (1/T^*)dx/dt - \kappa^* T_{x^*}^*/T^*$$

so that the heat balance boundary condition on $x^* = X^*$ becomes

$$\begin{aligned}\kappa^* \frac{\partial T^*}{\partial x^*} &= \left[\frac{-L\rho T^{*2}}{1 + L\rho T^*} \right] dx^*/dt^* \\ &= \frac{-L\rho dx^*/dt^*}{\Phi(T_f) [\Phi(T_f) + L\rho]} \quad \text{on } x^* = X^* .\end{aligned}\tag{22}$$

Thus, to summarize, under the reciprocal transformation (16), the moving boundary value problem (7)-(10) reduces, subject to the Sturm condition (17) to

$$\begin{aligned}\frac{\partial T^*}{\partial t^*} &= \kappa^* \frac{\partial^2 T^*}{\partial x^{*2}} , \\ \kappa^* \frac{\partial T^*}{\partial x^*} &= -UT^* \quad \text{on } x^* = \theta(t^*) - \theta(0) \\ \left. \begin{aligned}\kappa^* \frac{\partial T^*}{\partial x^*} &= - \frac{L\rho dx^*/dt^*}{\Phi(T_f) [\Phi(T_f) + L\rho]} \\ T^* &= \frac{1}{\Phi(T_f)}\end{aligned} \right\} \text{on } x^* = X^* \\ X^*(0) &= 0.\end{aligned}\tag{23}$$

Now,

$$dx = T^* dx^* + \kappa^* T_{x^*}^* dt^*$$

so that

$$\begin{aligned}\frac{\partial x}{\partial t^*} &= T^* \\ \frac{\partial x}{\partial t^*} &= \kappa^* T_{x^*}^* = \int_{X^*}^{x^*} \frac{\partial}{\partial x^*} (\kappa^* T_{x^*}^*) dx^* + \kappa^* T_{x^*}^* \Big|_{x^*=X^*} \\ &= \int_{X^*}^{x^*} \frac{\partial T^*}{\partial t^*} dx^* - \frac{L\rho dx^*/dt^*}{\Phi(T_f) [\Phi(T_f) + L\rho]}\end{aligned}$$

whence

$$x = \int_{X^*}^{x^*} T^* dx^* + X^* T^* \Big|_{x^*=X^*} - \frac{L\rho X^*}{\Phi(T_f) [\Phi(T_f) + L\rho]} .$$

Accordingly,

$$x(x^*, t^*) = \int_{X^*}^{x^*} T^* dx^* + X^* / [\Phi(T_f) + L\rho] . \quad (24)$$

Alternatively,

$$\begin{aligned} \frac{\partial x}{\partial t^*} &= T^* \\ \frac{\partial x}{\partial t^*} &= \int_{x^*=0}^{x^*} \frac{\partial}{\partial x^*} (\kappa^* T_{x^*}^*) dx^* + \kappa^* T_{x^*}^* \Big|_{x^*=0} \\ &= \int_{\theta(t^*)-\theta(0)}^{x^*} \frac{\partial T^*}{\partial t^*} dx^* - UT^* \Big|_{x^*=\theta(t^*)-\theta(0)} \end{aligned}$$

so that

$$x(x^*, t^*) = \int_{\theta(t^*)-\theta(0)}^{x^*} T^* dx^* . \quad (25)$$

Hence, if $T^*(x^*, t^*)$ is the solution of the reciprocal boundary value problem (23) then the solution of the original problem (7)-(10) is given parametrically by

$$T = \Phi^{-1}(1/T^*)$$

$$\begin{aligned} x &= \int_{\theta(t^*)-\theta(0)}^{x^*} T^* dx^* \\ t &= t^* \end{aligned} \quad (26)$$

while the evolution of the boundary $x = X$ is given in terms of

that of $x^* = X^*$ by the simple relation (22).

We now specialise our attention to the class of moving boundary problems (7)-(10) with

$$U(t) = U_0/\sqrt{t}, \quad X(t) = \sqrt{2\gamma t} \quad (27)$$

$$(28)$$

and introduce the similarity variable

$$\xi^* = x^*/\sqrt{2\gamma t^*} \quad (29)$$

into the reciprocal problem (23). Solutions are sought of the type

$$T^* = \phi^*(x^*/\sqrt{2\gamma t^*}) \quad (30)$$

so that

$$\gamma \xi^* \frac{d\phi^*}{d\xi^*} + \kappa^* \frac{d^2\phi^*}{d\xi^{*2}} = 0 \quad (31)$$

whence

$$\phi^* = A \operatorname{erf}[\sqrt{\frac{\gamma}{2\kappa^*}} \xi^*] + B. \quad (32)$$

The linear boundary conditions require that

$$\kappa^* \frac{d\phi^*}{d\xi^*} = -U_0 \sqrt{2\gamma} \phi^* \quad \text{on} \quad \xi^* = U_0 \sqrt{2/\gamma} \quad (33)$$

and

$$\phi^* = \frac{1}{\Phi(T_f)} \quad \text{on} \quad \xi^* = \Phi(T_f) + L\rho \quad (34)$$

whence, A and B are given by

$$A \sqrt{\frac{\kappa^*}{\pi}} e^{-U_0^2/\kappa^*} = -U_0 \left[A \operatorname{erf}\left[\frac{U_0}{\sqrt{\kappa^*}}\right] + B \right], \quad (35)$$

$$\text{Aerf}\left[\sqrt{\frac{\gamma}{2\kappa^*}} (\Phi(T_f) + L\rho)\right] + B = \frac{1}{\Phi(T_f)} . \quad (36)$$

The constant γ which determines the motion of the boundary $x = X(t)$ is obtained from the remaining boundary condition which provides the transcendental equation

$$A\sqrt{\frac{2\kappa^*}{\gamma\pi}} e^{-\frac{\gamma}{2\kappa^*} (\Phi(T_f) + L\rho)^2} = -L\rho/\Phi(T_f) . \quad (37)$$

The solution of the original boundary value problem is now given parametrically by the relations

$$T = \Phi^{-1}\left[\frac{1}{\text{Aerf}\left(\sqrt{\frac{\gamma}{2\kappa^*}} \xi^*\right) + B}\right] \quad (38)$$

$$\xi = \int_{U_0\sqrt{2/\gamma}}^{\xi^*} [\text{Aerf}\left(\sqrt{\frac{\gamma}{2\kappa^*}} \sigma\right) + B] d\sigma . \quad (39)$$

It is noted that the above class of exact solutions has the property that $T = \text{constant} = T_0$ on the boundary $x = 0$ where

$$\Phi(T_0) = \frac{1}{\text{Aerf}[U_0/\sqrt{\kappa^*}] + B} . \quad (40)$$

It is emphasised that the above analysis is valid for any member of the class of materials given by (17). It represents an extension to nonlinear heat conduction of the classical Neumann solution.

IV. TWO PHASE STEFAN PROBLEMS IN NONLINEAR HEAT CONDUCTION. APPLICATION OF RECIPROCAL TRANSFORMATIONS. The two-phase Stefan problem to be considered is for a semi-infinite

region $x > 0$ with phase change temperature T_f . It is required to determine the evolution of the moving phase separation boundary $x = X(t)$ and temperature distribution

$$T(x, t) = \begin{cases} T_2(x, t) > T_f & 0 < x < X(t) \\ T_1(x, t) < T_f & X(t) < x < \infty \end{cases} \quad (41)$$

where

$$\rho c_{p1}(T_1) \frac{\partial T_1}{\partial t} = \frac{\partial}{\partial x} [\kappa_1(T_1) \frac{\partial T_1}{\partial x}], \quad X(t) < x < \infty \quad (42)$$

$$\left. \begin{aligned} \kappa_1(T_1) \frac{\partial T_1}{\partial x} - \kappa_2(T_2) \frac{\partial T_2}{\partial x} &= L \rho \dot{X} \\ T_1 &= T_2 = T_f \end{aligned} \right\} \text{ on } x = X(t) \quad (43)$$

$$\rho c_{p2}(T_2) \frac{\partial T_2}{\partial t} = \frac{\partial}{\partial x} [\kappa_2(T_2) \frac{\partial T_2}{\partial x}], \quad 0 < x < X(t) \quad (44)$$

$$\kappa_2(T_2) \frac{\partial T_2}{\partial x} = U(t) \quad \text{on } x = 0, t > 0 \quad (45)$$

together with the initial conditions

$$X(0) = 0 \quad (46)$$

$$T_1(x, 0) = V_0 < T_f, \quad x > 0 \quad (47)$$

In the above, the $T_i(x, t)$, $c_{pi}(T_i)$, $\kappa_i(T_i)$ $i = 1, 2$ represent, in turn, the temperature distribution, specific heat and thermal conductivity in the two phases. The subscripts $i = 1, 2$ refer to the new and original phases respectively. In this problem a melting process is envisaged in

which phase 1 is solid and phase 2 is liquid. L denotes the latent heat of fusion of the medium. Here $U(t)$ denotes the prescribed flux on the boundary $x = 0$ while V_0 represents the initial temperature of the medium. It is noted that the analogous two-phase problem in linear heat conduction has been recently investigated by Tarzia [6]. As in that work and in the single phase problem, attention is restricted to the class of moving boundary problems with

$$U(t) = U_0/\sqrt{t}, \quad X(t) = \sqrt{2\gamma t}. \quad (48)$$

$$(49)$$

If we now set

$$\bar{T}_i = \phi_i(T_i) = \int_{T_{0i}}^{T_i} S_i(\sigma) d\sigma, \quad S_i = \rho c_{pi}(T_i), \quad i = 1, 2 \quad (50)$$

then (42) and (44) yield

$$\frac{\partial \bar{T}_i}{\partial t} - \frac{\partial}{\partial x} \left(\frac{\kappa_i}{\phi_i} \frac{\partial \bar{T}_i}{\partial x} \right) = 0 \quad i = 1, 2. \quad (51)$$

Our attention is henceforth confined to materials for which the Sturm conditions

$$\bar{\kappa}_i \phi_i' / \phi_i^2 = \kappa_i(T_i) \quad i = 1, 2 \quad (52)$$

apply, where $\bar{\kappa}_i$, $i = 1, 2$ are positive constants.

Use of the conditions (52) in (51) reduces the heat conduction equations in the two phases to the form

$$\frac{\partial \bar{T}_i}{\partial t} - \bar{\kappa}_i \frac{\partial}{\partial x} \left(\frac{1}{\bar{T}_i^2} \frac{\partial \bar{T}_i}{\partial x} \right) = 0, \quad i = 1, 2. \quad (53)$$

The similarity variable

$$\xi = x/\sqrt{2\gamma t} \quad (54)$$

is now introduced and solutions of (53) are sought in the form

$$\bar{T}_i = \phi_i(x/\sqrt{2\gamma t}) \quad i = 1, 2 \quad (55)$$

whence (53) yields

$$\gamma \xi \frac{d\phi_i}{d\xi} + \bar{\kappa}_i \frac{d}{d\xi} \left(\frac{1}{2} \frac{d\phi_i}{d\xi} \right) = 0 \quad i = 1, 2. \quad (56)$$

Under the reciprocal transformation

$$\left. \begin{aligned} d\xi &= \phi_i^* d\xi_i^* \\ \phi_i^* &= \frac{1}{\phi_i} \end{aligned} \right\} R \quad (R^2 = I) \quad (57)$$

(56) produces the linear canonical form

$$\bar{\kappa}_i \frac{d^2 \phi_i^*}{d\xi_i^{*2}} + \gamma \xi_i^* \frac{d\phi_i^*}{d\xi_i^*} = 0 \quad i = 1, 2 \quad (58)$$

with solution

$$\phi_i^* = A_i \operatorname{erf} \left[\sqrt{\frac{\gamma}{2\bar{\kappa}_i}} \xi_i^* \right] + B_i \quad i = 1, 2. \quad (59)$$

The four conditions

$$\begin{aligned} T_1 &= T_2 = T_f \quad \text{on} \quad x = X(t) \\ \bar{\kappa}_2(T_2) \frac{\partial T_2}{\partial x} &= \frac{U_0}{\sqrt{t}} \quad \text{on} \quad x = 0, t > 0 \\ T_1(x, 0) &= V_0 \end{aligned}$$

produce, in turn, four equations which determine the A_i , B_i

$i = 1, 2$ namely

$$A_1 \operatorname{erf} \left[\sqrt{\frac{\gamma}{2\bar{\kappa}_1}} \lambda_1 \right] + B_1 = \frac{1}{\phi_1(T_f)} , \quad (60)$$

$$A_2 \operatorname{erf} \left[\sqrt{\frac{\gamma}{2\bar{\kappa}_2}} \lambda_2 \right] + B_2 = \frac{1}{\phi_2(T_f)} , \quad (61)$$

$$A_2 \sqrt{\frac{\bar{\kappa}_2}{\pi}} \exp(-U_0^2/\bar{\kappa}_2) = -U_0 [A_2 \operatorname{erf}(U_0/\sqrt{\bar{\kappa}_2}) + B_2] \quad (62)$$

$$A_1 + B_1 = \frac{1}{\phi_1(V_0)} , \quad (63)$$

where

$$\lambda_1 = \xi_1^* \Big|_{\xi=1} , \quad \lambda_2 = \xi_2^* \Big|_{\xi=1} . \quad (64)$$

$$(65)$$

The interface condition

$$\kappa_1(T_1) \frac{\partial T_1}{\partial x} - \kappa_2(T_2) \frac{\partial T_2}{\partial x} = L\rho \dot{X} \quad \text{on} \quad x = X(t)$$

yields

$$-\frac{\bar{\kappa}_1}{\phi_1^*} \frac{d\phi_1^*}{d\xi_1^*} + \frac{\bar{\kappa}_2}{\phi_2^*} \frac{d\phi_2^*}{d\xi_2^*} = L\rho\gamma \quad \text{on} \quad \xi = 1$$

whence

$$\begin{aligned} -A_1 \phi_1(T_f) \sqrt{\frac{2\bar{\kappa}_1}{\gamma\pi}} \exp(-\gamma\lambda_1^2/2\bar{\kappa}_1) \\ + A_2 \phi_2(T_f) \sqrt{\frac{2\bar{\kappa}_2}{\gamma\pi}} \exp(-\gamma\lambda_2^2/2\bar{\kappa}_2) = L\rho . \end{aligned} \quad (66)$$

The latter provides a transcendental equation for the constant γ which determines the motion of the moving boundary $x = X(t) = \sqrt{2\gamma t}$.

The required temperature distributions T_1 and T_2 are

given parametrically by

$$\left. \begin{aligned} T_1 &= \phi_1^{-1} \{ A_1 \operatorname{erf} [\sqrt{\frac{\gamma}{2\kappa_1}} \xi_1^*] + B_1 \}^{-1} \\ \xi &= \int_{\lambda_1}^{\xi_1^*} \{ A_1 \operatorname{erf} [\sqrt{\frac{\gamma}{2\kappa_1}} \sigma] + B_1 \} d\sigma + 1 \end{aligned} \right\} \quad (67)$$

and

$$\left. \begin{aligned} T_2 &= \phi_2^{-1} \{ A_2 \operatorname{erf} [\sqrt{\frac{\gamma}{2\kappa_2}} \xi_2^*] + B_2 \}^{-1} \\ \xi &= \int_{U_0 \sqrt{2/\gamma}}^{\xi_2^*} \{ A_2 \operatorname{erf} [\sqrt{\frac{\gamma}{2\kappa_2}} \sigma] + B_2 \} d\sigma . \end{aligned} \right\} \quad (68)$$

The quantities λ_1 and λ_2 are given by the relations

$$\lambda_1 - \lambda_2 = L\rho + \phi_1(T_f) - \phi_2(T_f) \quad (69)$$

together with

$$1 = \int_{U_0 \sqrt{2/\gamma}}^{\lambda_2} A_2 \operatorname{erf} [\sqrt{\frac{\gamma}{2\kappa_2}} \sigma] d\sigma + B_2 [\lambda_2 - U_0 \sqrt{2/\gamma}] \quad (70)$$

References

- [1] Storm, M. L., J. Appl. Phys. 22, 940 (1951).
- [2] Rogers, C., J. Phys. A: Math. Gen. 16, L 493 (1983).
- [3] Rogers, C., J. Math. Phys. 26, 393 (1985).
- [4] Rubinstein, L. I., The Stefan Problem, Trans. Math. Monographs 27 (Providence: Amer. Math. Soc.) (1971).
- [5] Kingston, J. G. and Rogers, C., Phys. Lett. A 92, 261 (1982).
- [6] Tarzia, D. A., Quart. Appl. Math. 49, 491 (1982).

ANALYSIS OF FLUID EQUATIONS BY GROUP METHODS

W. F. Ames¹ and M. C. Nucci^{2,3}
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT. Using the machinery of Lie group analysis several equations arising in fluid mechanics are studied. In particular, the Burgers' equation the KdV equation, the Hopf equation, the two dimensional KdV equation and the Lin-Tsien equation are analyzed. In all cases the particular group includes arbitrary functions of time which permit the transformation of time dependent equations into the corresponding time independent ones. Infinitely many time dependent solutions are associated with each steady solution. Some solutions are constructed.

I. INTRODUCTION. Perhaps the most widely applicable method for determining analytic solutions of partial differential equations utilizes the underlying (Lie) group structure. The mathematical foundations for the determination of the full group for a system of differential equations can be found in Ames [1], Bluman and Cole [2], and the general theory is found in Ovsiannikov [3]. The determination of the full group requires extremely lengthy calculations. Detailed calculations can be found in Ames [1], Ovsiannikov [3] and for the Navier Stokes equations in Boisvert [4] (see also Boisvert, et al. [5]). Algebraic programming packages for determining these groups have been developed by Schwarz using REDUCE [9], by Roseneau and Schwarzmeier using MACSYMA [10] and CINO in Russia (see Ovsiannikov [3], p. 57). These programs, while very versatile, have difficulties in incorporating arbitrary functions where they arise in the Lie algebra. These arbitrary functions play a fundamental role in the sequel.

In Boisvert, et al. [5] the full Lie group leaving the Navier-Stokes equations invariant,

$$u_t + uu_x + vu_y + wu_z = -p_x + \mu \nabla^2 u, \quad (1.1)$$

¹Research supported by U.S. Army Grant DAAG-29-84-K-0083.

²Permanent address: Dipartimento di Matematica, Università di Perugia, 06100 Perugia, Italy

³Research supported by a NATO-CNR fellowship.

$$v_t + uv_x + vv_y + wv_z = -p_y + \mu \nabla^2 v, \quad (1.2)$$

$$w_t + uw_x + vw_y + ww_z = -p_z + \mu \nabla^2 w, \quad (1.3)$$

$$u_x + v_y + w_z = 0, \quad (1.4)$$

is determined. In the spirit of Lie it is desired to find infinitesimal transformations of the form

$$\begin{aligned} t' &= t + \epsilon T(t, x, y, z, u, v, w, p) + O(\epsilon^2), \\ x' &= x + \epsilon X(t, x, y, z, u, v, w, p) + O(\epsilon^2), \\ y' &= y + \epsilon Y(t, x, y, z, u, v, w, p) + O(\epsilon^2), \\ z' &= z + \epsilon Z(t, x, y, z, u, v, w, p) + O(\epsilon^2), \\ u' &= u + \epsilon U(t, x, y, z, u, v, w, p) + O(\epsilon^2), \\ v' &= v + \epsilon V(t, x, y, z, u, v, w, p) + O(\epsilon^2), \\ w' &= w + \epsilon W(t, x, y, z, u, v, w, p) + O(\epsilon^2), \\ p' &= p + \epsilon P(t, x, y, z, u, v, w, p) + O(\epsilon^2), \end{aligned} \quad (1.5)$$

which leave (1.1-1.4) invariant. System (1.5) leaves (1.1-1.4) invariant if and only if (u', v', w', p') is a solution of (1.1'-1.4') whenever (u, v, w, p) is a solution to (1.1-1.4). By (1.1'-1.4') is meant the same equations in the primed variables. By extensive analysis it is found that the full Lie group leaving (1.1-1.4) invariant is given by (1.5) with

$$T = \alpha + 2\beta t, \quad (1.6)$$

$$X = \beta x - \gamma y - \lambda z + f(t), \quad (1.7)$$

$$Y = \beta y + \gamma x - \sigma z + g(t), \quad (1.8)$$

$$Z = \beta z + \lambda x + \sigma y + h(t), \quad (1.9)$$

$$U = -\beta u - \gamma v - \lambda w + f'(t) \quad (1.10)$$

$$V = -\beta v + \gamma u - \sigma w + g'(t) \quad (1.11)$$

$$W = -\beta w + \lambda u + \sigma v + h'(t) \quad (1.12)$$

$$P = -2\beta p + j(t) - xf''(t) - yg''(t) - zh''(t) \quad (1.13)$$

where $\alpha, \beta, \gamma, \lambda$ and σ are five arbitrary parameters and $f(t)$, $g(t)$, $h(t)$ and $j(t)$ are arbitrary, sufficiently smooth functions of t .

The arbitrary functions in (1.7-1.9) permit equations (1.1-1.4) to be transformed into their time-independent form. Thus any solution of the steady equations generates an infinity of time dependent solutions. This idea was exploited in Boisvert, et al. [5] and Nucci [6].

II. BURGER'S EQUATION. For the Burgers' equation

$$u_t + uu_x = \mu u_{xx} \quad (2.1)$$

the full two parameter group (α, β) with one arbitrary function $(f(t))$ is

$$T = \alpha + 2\beta t, \quad X = \beta x + f(t), \quad U = -\beta u + f'(t) \quad (2.2)$$

With $\alpha = 1$, $\beta = 0$ the subgroup

$$T = 1, \quad X = f(t), \quad U = f'(t)$$

has the generator

$$QI = \frac{\partial I}{\partial t} + f(t) \frac{\partial I}{\partial x} + f'(t) \frac{\partial I}{\partial u} = 0. \quad (2.3)$$

From the Lagrange equations of (2.2) we have

$$\begin{aligned} \bar{u} &= u - f(t) \\ \bar{x} &= x - F(T), \end{aligned} \quad (2.4)$$

where $F' = f$. When this transformation is applied to (2.1) there results

$$\bar{u}\bar{u}_{\bar{x}} = \mu \bar{u}_{\bar{x}\bar{x}}, \quad (2.5)$$

that is the steady equation. One integration gives the integrable Riccati equation

$$U' + U^2 = C, \quad (2.6)$$

where $u = -2\mu U$. Finally setting $U = \psi'/\psi$ (2.6) becomes

$$\psi'' - C\psi = 0.$$

For $C = \alpha^2 > 0$ the solution of (2.1) is

$$u(x,t) = -2\mu\alpha\{1 - \text{Rexp}[-2\alpha(x-F(t))]\} / \{1 + \text{Rexp}[-2\alpha(x-F(t))]\} + f(t).$$

For $C = -\alpha^2 < 0$ the the solution for (2.1) is

$$u(x,t) = -2\mu\alpha \frac{\cos[\alpha(x-F(t))] - R \sin[\alpha(x-F(t))]}{\sin[\alpha(x-F(t))] + R \cos[\alpha(x-F(t))]} + f(t)$$

where R is another arbitrary constant. If $C = 0$ the solution is

$$u(x,t) = \frac{2\mu}{2\mu R - (x-F(t))} + f(t).$$

III. THE KORTEWEG DE VRIES EQUATION. Under the transformation (2.4) the KdV equation $u_t + uu_x = u_{xxx}$ becomes

$$\bar{u}\bar{u}_x = \bar{u}_{xxx}$$

which is integrable in terms of elliptic functions since one integration gives

$$U_{xx} - U^2 = C,$$

when $u = 2U$.

IV. THE EQUATION $u_t + uu_x = [\phi(u_x)u_x]_x$. Again the action of (2.4) transforms the equation of the title into

$$\bar{u}\bar{u}_x = [\phi(\bar{u}_x)\bar{u}_x]_x.$$

V. TWO DIMENSIONAL K-dV EQUATION. The full group for the two dimensional K-dV equation (Rogers and Chadwick [7])

$$u_{xxxx} = -u_{xt} - \alpha u_{yy} - 6u_x^2 - 6uu_{xx} \quad (5.1)$$

is calculated, using the notation of (1.5) for t, x, y , and u , to be

$$T = f(t)$$

$$X = f'(t)x/3 - f''(t)y^2/6\alpha - g'(t)y/2\alpha + h(t)$$

$$Y = 2f'(t)y/3 + g(t) \quad (5.2)$$

$$U = -2f'(t)u/3 + f''(t)x/18 - f'''(t)y^2/36\alpha - g''(t)y/12\alpha + h'(t)/6.$$

In (5.2) the functions $f(t), g(t)$ and $h(t)$ are arbitrary so

the Lie Algebra is infinite dimensional with the generator

$$Q = T \frac{\partial}{\partial t} + X \frac{\partial}{\partial x} + Y \frac{\partial}{\partial y} + U \frac{\partial}{\partial u}. \quad (5.3)$$

Moreover, the specific choice of the subgroup

$$\begin{aligned} T &= 1, & X &= -g'y/2\alpha + h(t), & Y &= g(t) \\ U &= -g''y/12\alpha + h'/6 \end{aligned} \quad (5.4)$$

gives rise to the characteristic variables

$$y - \bar{y} = G(t)$$

$$x - \bar{x} = -gy/2\alpha + \Omega(t) \quad (5.5)$$

$$u - \bar{u} = -g'y/12\alpha + g^2/24\alpha + h/6,$$

where $G' = g$ and $\Omega' = g^2/2\alpha + h$. Under (5.5) equation (5.1) becomes

$$\alpha \bar{u}_{\bar{y}\bar{y}} + 6(\bar{u}_{\bar{x}})^2 + 6\bar{u}\bar{u}_{\bar{x}\bar{x}} + \bar{u}_{\bar{x}\bar{x}\bar{x}\bar{x}} = 0, \quad (5.6)$$

that is the time independent equation. Each solution of (5.6) gives rise to a family of solutions involving three arbitrary functions of time, f , g and h , when \bar{u} , \bar{x} , and \bar{y} are replaced by their relations from (5.5). To illustrate this idea the full group is generated for equation (5.6) and some exact solutions are obtained in the next section.

VI. SOLUTIONS FOR EQUATION (5.6). The full group for equation (5.6) (we have dropped the bars), in the notation of (1.5) for x , y , and u is

$$\begin{aligned} X &= c_1 x + c_2 \\ Y &= 2c_1 y + c_3 \\ U &= -2c_1 u, \end{aligned} \quad (6.1)$$

with the three parameters c_1, c_2, c_3 . The equation for the invariant surface are obtained from (for $c_1 \neq 0$)

$$\begin{aligned} u &= F(\eta)/2c_1(2c_1 y + c_3) \\ \eta &= (c_1 x + c_2)/(2c_1 y + c_3)^{1/2}, \end{aligned} \quad (6.2)$$

where F satisfies the ordinary differential equation

$$c_1^3 F^{(iv)} + 3FF'' + \alpha c_1 \eta^2 F'' + 3(F')^2 + 7c_1 \alpha \eta F' + 8\alpha c_1 F = 0. \quad (6.3)$$

Two solutions of equation (6.3) are

$$F(\eta) = -4\alpha c_1 \eta^2 / 3 \quad (6.4)$$

and

$$F(\eta) = -4c_1^3 \eta^{-2}. \quad (6.5)$$

The solution of equation (5.1) resulting from (6.4) is

$$u = - \frac{2\alpha \{c_1 [x + \frac{g}{2\alpha} y - \Omega] + c_2\}^2}{3\{2c_1[y - G(t)] + c_3\}^2} - \frac{g'}{12\alpha} y + \frac{g^2}{24\alpha} + \frac{h}{6}$$

and that resulting from (6.5) is

$$u = - \frac{2c_1^2}{\{c_1 [x + \frac{g}{2\alpha} y - \Omega(t)] + c_2\}^2} - \frac{g'}{12\alpha} y + \frac{g^2}{24\alpha} + \frac{h}{6}$$

where $\Omega' = g^2/24\alpha + h$ and $G' = g$.

VII. AN INTERESTING SUBGROUP OF (5.2). With the choice $f(t) = t^3$, $g(t) = h(t) = 0$ in (5.2) the group becomes

$$\begin{aligned} T &= t^3, & X &= t^2 x - ty^2/2, & Y &= 2t^2 y \\ U &= -2t^2 u + tx/3 - y^2/6\alpha \end{aligned} \quad (7.1)$$

From (7.1) the equations for the invariant surface are found from

$$\eta = y/t^2, \quad \xi = x/t + y^2/2\alpha t^2 \quad (7.2)$$

and

$$u = \xi/6 - \eta^2 t^2/12\alpha + F(\eta, \xi)/t^2, \quad (7.3)$$

where $F(\eta, \xi)$ satisfies the equation

$$\alpha F_{\eta\eta} + (3F^2)_{\xi\xi} + F_{\xi\xi\xi\xi} = 0. \quad (7.4)$$

But, of course, this is equation (5.6) with $\eta = \bar{y}$, $\xi = \bar{x}$ and $F = \bar{u}$. Thus two solutions are available - i.e.,

$$u(x, t) = \frac{x}{6t} - \frac{2\alpha [c_1 xt + c_1 y^2/2\alpha + t^2 c_2]^2}{3t^2 [2c_1 y + t^2 c_3]^2}$$

and

$$u(x,t) = \frac{x}{6t} - \frac{2c_1^2 t^2}{(c_1 x t + c_1 y^2 / 2\alpha + t^2 c_2)^2}.$$

[8] VIII. THE LIN-TSIEN EQUATION. The Lin-Tsien equation

$$2\phi_{tx} + \phi_x \phi_{xx} - \phi_{yy} = 0, \quad (8.1)$$

where ϕ is a velocity potential, has been used to study dynamic transonic flow in two space dimensions (x and y). The full group is known (see Ovsiannikov [3, p. 388]) to be

$$\begin{aligned} X &= 4\alpha x/3 + g'(t)y + w(t) \\ Y &= \alpha y + g(t) \\ T &= 2\alpha t/3 + \beta \end{aligned} \quad (8.2)$$

$$\phi = 2\alpha\phi + 2y^3 g'''/3 + 2y^2 w'' + 2xyg'' + 2xw' + yr(t) + s(t).$$

Equations (8.2) contain two arbitrary constants, α and β , and four arbitrary functions $g(t)$, $w(t)$, $r(t)$ and $s(t)$. Consequently, the Lie Algebra is infinite dimensional. Once again these arbitrary functions will permit an infinite number of time dependent solutions to be generated from each steady state solution.

The invariants of the group that are constant in time (with $\alpha = 0$) are found as before to be

$$\begin{aligned} x - \bar{x} &= gy - \lambda, \quad \lambda = \int (g^2 + w) dt \\ y - \bar{y} &= Q, \quad Q = \int g(t) dt \end{aligned}$$

and

$$\begin{aligned} \phi - \bar{\phi} &= 2g'\bar{x}\bar{y} + 2g''\bar{y}^3/3 + 2w\bar{x} \\ &+ \int \{2g''[g\bar{y}^2 + 2gQ\bar{y} - \lambda\bar{y} + gQ^2 - \lambda Q] \\ &+ \frac{2}{3} g''' [3\bar{y}^2 Q + 3\bar{y}Q^2 + Q^3] + 2w'[\bar{y}g + gQ - \lambda] \\ &+ 2w''[2\bar{y}Q + Q^2] + r\bar{y} + s\} dt \end{aligned} \quad (8.3)$$

It is easy to show that $\bar{\phi}(x,y)$ satisfies the time independent equation

$$\bar{\phi}_{\bar{x}} \bar{\phi}_{\bar{x}\bar{x}} - \bar{\phi}_{\bar{y}\bar{y}} = 0 \quad (8.4)$$

which has been much studied. Given any solution of (8.4) it

follows that time dependent solutions (with four arbitrary functions of time g , w , r and s) are constructable -- thus

$$\phi(x,y,t) = \bar{\phi}(x - g(t)y + \lambda, y - Q)$$

+ (right hand side of the last equation in (8.3)).

References

1. W. F. Ames, Nonlinear Partial Differential Equations in Engineering, Vol. II, Chapter 2, Academic Press, New York, 1972.
2. G. W. Bluman and J. D. Cole, Similarity Methods for Differential Equations, Springer, New York, 1974.
3. L. V. Ovsiannikov, Group Analysis of Differential Equations (Russian Edition, NAUKA 1978); English translation edited by W. F. Ames, Academic Press, 1982.
4. R. E. Boisvert, Group Analysis of the Navier-Stokes Equations, Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, GA 30332, 1982.
5. R. E. Boisvert, W. F. Ames and U. N. Srivastava, Group Properties and New Solutions of Navier-Stokes Equations, Journal of Engineering Math. 17, 203 (1983).
6. M. C. Nucci, Group Analysis for MHD Equations, in press, Atti Sem. Mat. Fis. Univ. Modena 33 (1984).
7. C. Rogers and W. F. Shadwick, Bäcklund Transformations and Their Applications, Academic Press, N.Y., 1982.
8. K. Oswatitsch (Editor), Symposium Transonicum (Aachen, 1962), Springer-Verlag Berlin, 1964.
9. F. Schwarz. A Reduce Package for Determining Lie Symmetries of Ordinary and Partial Differential Equations, Comp. Physics Commun. 27, p. 179-186, 1982.
10. P. Roseneau and J. L. Schwarzmeier, Similarity Solutions of Systems of Partial Differential Equations Using MACSYMA, Courant Inst. of Math. Sci. Report No. C00-3077-160/MF-94, 1979.

F-P-S POINCARÉ-LIKE LINEARIZATION APPLIED TO SOLITON EQUATIONS

R.L. Anderson

Department of Physics and Astronomy
University of Georgia, 30602, U.S.A.

and

E. Taflin

Département de Physique Théorique
Université de Genève
CH-1211, Genève 4,
Switzerland

ABSTRACT. The program of F-P-S Poincaré-like linearization and the results of its application to soliton equations are described.

I. Introduction. The results announced in [1] strongly suggest that soliton equations are F-P-S Poincaré-like linearizable [5]. This was further supported by our work [2,3,4]. The purpose of this paper is to describe heuristically the F-P-S Poincaré-like linearization program (§II) and some of its implications (see last section of §II), as well as describe the results obtained to date from its application to soliton equations (§III).

One important reason for continued physical interest in soliton equations such as the Korteweg de Vries (KdV), nonlinear Schrödinger, Boussinesq, and the Benjamin-Ono (B.O.) equations is that they are the lowest order nonlinear equations in multiscale approximations to the equations of one-dimensional fluid flow for various physical configurations and certain initial data. They are of interest in their own right because they possess an infinite-dimensional Lie symmetry structure which can be used to generate hierarchies of constants of motion and classes of soliton solutions. They exhibit particle-like properties. Further, it is known that many soliton equations are particular cases of completely integrable infinite-dimensional Hamiltonian systems [6,7]. There are various effective approaches to studying such equations: e.g., IST Scheme [8,9,10,11]; Riemann problem for matrices [12]; Method of Prolongation Structures [13] (the reader is also referred to the presentation by Estabrook in these proceedings); Kac-Moody Lie algebraic methods [14], (see [14] for the references to the work of the Kyoto School.), loop groups [15], Riemann surfaces and theta functions [16]. What is the justification for the interest in all these approaches and, as we advocate, even one more? To date there does not exist an algorithmic method for determining whether a given equation is of the soliton type; rather all of the above provide methods for constructing soliton equations. In the case of the Method of Prolongation Structures it also provides a way for searching for structures such as symmetry, Lax pairs, Bäcklund transformations for a given equation, if they exist. We have approached this general area with two broad goals in mind. One is to realize the implications of the F-P-S Poincaré linearization program cited at the end of §II. This includes, of course, solving the evolution equation in question. The second goal is to use this program as a tool to try to determine algorithmically whether a given equation is a soliton equation or not. In §III we describe our results to date which show some progress towards these goals.

II. F-P-S Poincaré-like linearization. The program of the classification of nonlinear representations of Lie groups and Lie algebras of Flato-Pinczon-Simon [5,17] can be regarded as a generalization of Poincaré's program of transforming vector fields to normal forms. Here we shall briefly describe this transformation for the case when there exists an equivalent linear normal form in the neighborhood of the origin. Further, we shall confine our discussion to the following formulation which is sufficient for introducing the general ideas. Consider a nonlinear analytic vector field T with Taylor expansion about the origin of the form $T = \sum T^n$ where T^n is the n^{th} order term, which

$$n \geq 1$$

describes a nonlinear ordinary differential equation (ODE) of the form $du/dt = T(u)$, $t \in \mathbb{R}$, $u \in \mathbb{R}^n$. We look for ways to transform this vector field (this ODE) to a simpler form. That is the general problem. The problem we are interested in here is to determine whether or not this vector field (ODE) can be transformed to its linear part through the use of power series substitutions. Explicitly, we seek to determine whether or not there exists an analytic map A , acting on some neighborhood of the origin, such that

$$DA \cdot [T] = T^1 \circ A, \quad (\text{II.1})$$

where $DA[T]$ is the Fréchet derivative of A along the direction T and \circ denotes composition of maps. Such a map $A: u \rightarrow v$ takes solutions of $du/dt = T(u)$ into solutions of $dv/dt = T^1(v)$. In the Taylor expansions T^n and A^n can be chosen to be symmetric n -linear maps. Therefore substitution of the formal power series representations for T and A into (II.1) yields the following equations

$$[T^1, A^n]_* = \sum_{1 \leq p \leq n-1, 0 \leq q \leq p-1} AP(I_q \otimes T^{n-p+1} \otimes I_{p-q-1}) \sigma_n, \quad (\text{II.2})$$

where $[T^1, A^n]_* = DT^1 \cdot [A^n] - DA^n[T^1]$, $I_q = q$ -fold tensor product of the identity operator, and σ_n is the symmetrization operator on the n -fold tensor product. (The tensor product appears because it is possible and useful to consider linear maps rather than n -linear maps at each order). Such a map is called a linearization map.

An analogous discussion for a map B which takes T^1 into T under the same assumptions yields the following equations

$$[B^n, T^1]_* = \sum_{1 \leq p \leq n-1, i_1 + \dots + i_p = n-1} TP \circ (B^{i_1} \otimes \dots \otimes B^{i_p}) \sigma_n. \quad (\text{II.3})$$

The map B is called an inverse linearization map and it is this map which underlies the usual perturbation methods. In fact, it is somewhat surprising that this approach of Poincaré to perturbation computations is not more commonly used. This is what is termed Poincaré linearization and we now proceed to set it in a group theoretical context and then describe its generalization à la Flato-Pinczon-Simon to what is termed F-P-S Poincaré-like linearization.

Equation (II.2) can be given the following group theoretical interpretation. Consider the local group action obtained by integrating the vector field $T(T^1)$ to $U_t(U_t^1)$, where t is the group parameter. Then (II.2) is the Lie algebraic version of the problem of the equivalence of the

one parameter group U_t to the linear group U_t^1 , i.e. does there exist an analytic map A intertwining U_t and U_t^1 ,

$$A \circ U_t = U_t^1 \circ A \quad . \quad (II.4)$$

We remind the reader that U_t and A appearing in equation (II.4) are nonlinear maps.

It is this part of the program Poincaré linearization which Flato-Pinczon-Simon generalized to classify the actions of nonlinear representations of Lie groups and Lie algebras on, in general, infinite-dimensional spaces. In what is termed F-P-S Poincaré-like linearization a given nonlinear evolution equation on a given space of initial conditions defines a one-parameter time-translation subgroup of a (larger) covariance group (such as the Galilean group or Poincaré group) of the equation in question. A choice of a covariance (symmetry) group is naturally suggested by the physical context in which the equation arises. The existence of an explicit invertible analytic linearization map A from a space of initial conditions for the nonlinear equation to one for the linear equation which takes the nonlinear representation of the covariance algebra into its linear part is investigated by analyzing the linearization algorithm (II.2) and the inverse linearization algorithm (II.3). This turns out to be a problem of cohomology, the details of which we omit here. We shall illustrate F-P-S Poincaré-like linearization with an application to the B.O. equation.

The Benjamin-Ono (B.O.) equation can be written in the form

$$du(t)/dt = T_0(u(t)), \quad u(t) \in E, t \in \mathbb{R}, \quad (II.5)$$

where T_0 is one generator of the nonlinear representation $\mathbb{R}^2 \ni (a,b) \rightarrow aT_0 + bT_1$ of the commutative space-time Lie algebra \mathbb{R}^2 . Here for

$$w \in E: T_0 = T_0^1 + T_0^2, \quad T_1 = T_1^1, \quad T_0^1(w) = -\partial^2 Hw, \quad T_0^2(w) = -\partial(w)^2, \\ T_1^1(w) = \partial w, \quad (\partial w)(x) = dw(x)/dx \quad \text{and} \quad (Hw)(x) = \text{P.V.} \int dy w(y)(y-x)^{-1} \pi^{-1}$$

is the Hilbert transform. E stands for various spaces, which in our work are all subspaces of Sobolev spaces $W^{n,2}$, the space of functions which are in $L^2(\mathbb{R})$ together with all their derivatives, up to order n . Its linear part is given by

$$dv(t)/dt = T_0^1(v(t)), \quad v(t) \in E, t \in \mathbb{R}. \quad (II.6)$$

Suppose for a moment that the evolution operator U_t (resp. U_t^1) associated with the Equation (II.5) (resp. (II.6)) exists, for each time t , on a space E of initial condition, i.e., if u_0 (resp. v_0) is an initial condition for equation (II.5) (resp. (II.6)) at $t = 0$, then the solution of Equation (II.5) (resp. (II.6)) is

$$u(t) = U_t(u_0) \quad (\text{resp. } v(t) = U_t^1(v_0)) \quad . \quad (II.7)$$

Equation (II.7) is the one-parameter time-translation group, with group parameter t , defined by the B.O. (linear B.O.) equation. It is sufficient for our purpose to consider this group as a one-parameter subgroup of a larger covar-

$$[U_{(t,a)}(u_0)](x) = [U_t(u_0)](x+a) \text{ and } [U_{(t,a)}^1(v_0)](x) = [U_t^1(v_0)](x+a) .$$

The map, from \mathbb{R}^2 into functions on E ,

$$(t,a) \longrightarrow U_{(t,a)} \text{ (resp. } U_{(t,a)}^1 \text{)} \quad (\text{II.8})$$

defines a nonlinear (resp. linear) representation of the commutative space-time covariance (symmetry) group \mathbb{R}^2 on E , i.e.,

$$U_{(t,a)} \circ U_{(t',a')} = U_{(t+t',a+a')} \quad (\text{resp. } U_{(t,a)}^1 \circ U_{(t',a')}^1 = U_{(t+t',a+a')}^1) .$$

The solution of (II.2) and (II.3) for the generators T_0 and T_1 of the group \mathbb{R}^2 , subject to the condition $B \circ A = \text{id}_E$, yields an invertible linearization map A with $B = A^{-1}$ which linearizes (II.8).

The implications of the existence of a solution B of Equations (II.3) are several. This solves one of the most fundamental problems about an evolution equation, namely, the initial value problem for the nonlinear equation on the set of initial conditions $\{u_0 \in E | u_0 = B(v_0)\}$ for some $v_0 \in E$ by

$$u(t) = B(v(t)), \quad u_0 = B(v_0) \quad (\text{II.9})$$

where $v(t)$ is the solution of the linear part with initial condition v_0 . Also other questions for the nonlinear evolution equation, such as the existence of superposition principles, hierarchies of 'higher order' evolution equations (or symmetries), and infinite sequence of constants of motion are reduced to the corresponding questions for the linear equation when A and B exist. We want to stress that we have not, in this heuristic discussion, considered the mathematically important and technical question of domains for A and B . Some of these points, as well as the B.O. equation itself, are discussed in more detail in §III and in the papers referenced there.

III. Description of results. It was known, since the discovery of the IST scheme for the KdV equation [8], how to construct solutions of the KdV equation from a certain class of solutions of the linear part. However, it was not clear how the Cauchy problem for the KdV equation on a given space of initial conditions reduces to one for its linear part. In [1] it was first shown how this reduction takes place by showing that on the space of initial conditions

$$S_b(\mathbb{R}) = \{f \in C^\infty(\mathbb{R}) | \|f\|_N = \sup_{\substack{x \in]-\infty, N] \\ 0 \leq k \leq N}} |(1+|x|)^{N_k} f(x)| < \infty, N = 0, 1, \dots\}$$

the Cauchy problem for the KdV equation can be solved entirely by the F-P-S Poincaré-like linearization program described in §II. Specifically, on the space of initial data S_b , a translation-invariant invertible analytic linearization map $A : S_b \rightarrow A[S_b] \subset S_b$ was concretely realized. (Technically in this case A is unique and converges to an entire analytic function on S_b , and is invertible.) The problem of the existence of global (in time) solutions and the existence of solutions which are not global (in time) was solved for the space S_b . In particular, as was earlier established in [18], the existence of global solutions for the KdV equation for all initial

conditions in $S(\mathbb{R})$, the Schwartz space of rapidly decreasing function on the real line \mathbb{R} , was established. Here the fact that $S(\mathbb{R}) \subset S_b(\mathbb{R})$ was used, as well as a separate fact [18], namely, $A[S(\mathbb{R})]$ is invariant under the linear evolution. A nonlinear superposition principle for the KdV equation was also obtained from the result that $A[S_b]$ is convex.

This program was next applied to Burgers equation. Although Burgers equation is not a soliton equation it shares many properties with soliton equations and is an important nonsoliton equation. Because of the existence of the Cole-Hopf transformation, it was well known that Burgers equation linearizes to its linear part. However, the application of the F-P-S Poincaré-like linearization program to Burgers equation led to new results concerning its constants of motion and Hamiltonian structure. For completeness we shall discuss the nature of the linearization map obtained in [19] as well as the other results.

In [19] Burgers equation was linearized on the space of initial condition $S(\mathbb{R})$. Specifically, a translation-invariant invertible analytic linearization map $A: S \rightarrow A[S] \subset S$ was concretely realized. (Technically A converges to an entire analytic function on S . The inverse $A^{-1}: A[S] \rightarrow S$ is analytic on the image $A[S]$). A^{-1} is close to the Cole-Hopf-transformation. It was shown that $A[S]$ is invariant under the linear evolution, hence the existence of global solutions for the Burgers equation for all initial conditions in $S(\mathbb{R})$ (or $L^1(\mathbb{R})$) was established. Further, it was established that Burgers equation has an infinity of constants of motion. This fact was not commonly thought to be true. It was also shown that $A[S]$ is a convex subset of S , hence the existence of a nonlinear superposition principle for Burgers equation was established. It was shown that the commutative Lie algebra of higher-order Burgers equations is the image of the positive powers of the translation operator under A^{-1} . Thus the enveloping algebra of the translation operator for the linear part is the source of the Burgers hierarchy. It was also shown that while Burgers' equation is dissipative it can be defined by a completely integrable Hamiltonian system.

The above result on the connection between the hierarchy of higher-order Burgers equations and the powers of the translation operator generalizes and the analogous result for the KdV equation was established earlier in [18] by analyzing the direct and inverse scattering maps in the IST scheme. In this latter work the question of determining a complete set of constants of motion for the KdV was solved. Specifically, it was shown that the Lie algebra spanned by certain polynomials in the generators of a solvable Lie algebra of dimension three for the linear part underlies each constant of motion for the KdV equation.

In [3] it was shown, by using examples, that the IST scheme, Hirota τ_N formalism, and the Kac-Moody constructions of the Kyoto School all yield the same inverse linearization operator in the sense described in §II. Hence, they solve the Cauchy problem for the same set of initial conditions. Specifically, in the case of a nonlinear evolution equation solvable by the IST scheme, iteration of the associated Gelfand-Levitan-Marchenko (G-L-M) equation yields the Taylor series representation of an inverse linearization map. For example, following the approach in [1], formulas (3.11), (4.02), and (4.14) of Rosales [20] can be given a rigorous mathematical meaning and hence represent analytic solutions to the cohomological equations for A^{-1} ; for the KdV, MKdV sine-Gordon, and nonlinear Schrödinger equations, respectively. The Kadomstev-Petviashvili (K-P) equation turns out to be another example. It was shown in the case of Hirota's τ_N formalism (or dependent variable substitution method) that it contains the essential

information for the construction of inverse linearization maps. This is illustrated for the K-P equation. To go from the N-soliton result to A^{-1} , one simply writes the N-soliton solution in terms of N solutions of the linear part of the equation. Then a generalization of this yields A^{-1} for more general initial conditions. Since the Kac-Moody construction of the Kyoto School in general subsumes the results obtained via Hirota's τ_N formalism a similar construction yields A^{-1} for the whole hierarchy. This is illustrated for a sine-Gordon hierarchy. (The latter example technically uses a generalization of the Kyoto School construction. This generalization is under separate investigation.) Given the central role the K-P hierarchy, its variations, and reductions, play in the extensive results of the Kyoto School, the class of soliton equations which are F-P-S Poincaré-like linearizable is quite large.

In [2] the cohomological equations for an inverse linearization map for the Benjamin-Ono equation were explicitly solved. A domain of convergence for this B (B_C in the language of [4]), which is invariant under the linear evolution, was identified. The resulting global solitonless solutions for the B.O. equation are for each time t decreasing at least as x^{-1} but not faster than x^{-2} in the space variable x for nontrivial initial data. Beyond its physical significance, this soliton equation was and still is extremely important in its own right because it has a dispersion relation $\omega(p) = |p|p$, $p \in \mathbb{R}$ for its linear part which is not analytic. (This dispersion relation has a certain similarity to the dispersion relation for the linear wave equation.)

The above analysis of the B.O. equation was continued in [4]. There, we extended the map B_C found in [2] to include solitons. Specifically, we have explicitly solved the Cauchy problem for certain initial data close to pure n-soliton data (Th 2.3, 3.1, [4]). This last result followed from an analysis of the cohomological equations for the linearization maps A . These latter equations were explicitly solved for a space-time covariant map A_C . The map A_C was shown to annihilate pure multi-solitons. Therefore in order to treat initial conditions which include multi-soliton components, an analysis of an associated linear problem which was constructed from a special property of A_C , which we loosely call its recursivity, led to the deformation of the map A_C into an analytic linearization map A which exists on a well-defined space of initial conditions (Th. 2.3 [4]) and is one-to-one on pure multi-solitons. The main technical problem remaining for solving explicitly the Cauchy problem for all initial conditions in any one of the spaces appearing in our analysis, is to characterize the elements in the image of A . A comparison of these results with the work of Fokas and Ablowitz [21] was made in [4]. Further comparison is actively being pursued by one of the authors (RLA) with the authors of [21].

An interesting result concerns the property we referred to in the preceding paragraph as the recursivity of the linearization (inverse linearization) maps. While we do not have a precise definition of what we mean by recursivity, computationally we identify this property in a given case by the simplification of the division problem which appears in the solution of the cohomological equations. While in all cases A^n is given in terms of all the A^j , $j < n$, and similarly for B^n it happens in these cases that they are given by iterating a few elementary operations. The recursivity of A_C leads naturally in the case of the B.O. equation to the identification of the Fourier transform of A_C with the distorted Fourier transform [22] associated

with a self adjoint operator which is the x-member of a Lax pair. This leads to a Lax pair which in the case of the positive frequency part is a particular case of the Lax-pair of Bock and Kruskal [23]. (This also leads directly to an explicit construction of an infinite hierarchy of constants of motion.) The recursivity of B_c leads to a Gelfand-Levitan-like equation.

With respect to the latter work on the B.O. equation, it is perhaps worthwhile to end by pointing out that the scattering operator for the Lax pair we constructed is a constant of motion, therefore our solution for the B.O. equation is a non-IST one. Further, the perturbation approach employed by Rosales [20] contains an ansatz which is implicitly a particular case of the algebraic part (i.e. no topological considerations) of space translation invariant formal F-P-S Poincaré-like inverse linearization. In light of our result for the B.O. equation, namely A and B are not space-translationally invariant. His approach will not yield without conceptual modification the solitonic part of B for the B.O. equation. Another much simpler example of a formally linearizable equation (with a covariance group reduced to the time-translations), which is not treatable by Rosales' approach, is

$$\frac{\partial}{\partial t} u(t,x) = \frac{\partial}{\partial x} u(t,x) + (u(t,x))^2.$$
 In this context, we remind the reader that the general F-P-S problem is the classification of nonlinear representations of Lie groups into normal forms where linearizable representations belong to the trivial equivalence class. The F-P-S Poincaré-like linearizability of a given evolution equations therefore depends critically on the covariance group considered as illustrated by our last example.

REFERENCES

- [1] Taflin, E., "Analytic Linearization of the Korteweg-deVries Equation," *Pac. J. Math.* 108, 203(1983).
- [2] Anderson, R.L. and Taflin, E., "Explicit Nonsoliton Solutions of the Benjamin-Ono Equation", *Lett. Math. Phys.* 7, 243(1983).
- [3] Anderson, R.L. and Taflin, E., "Linearization - a unified approach," pp. 19-23, in *Group Theoretical Methods in Physics*, Edited by G. Denardo G.Ghirardi, and T. Weber, *Lecture Notes in Physics*, 201(1984), Springer-Verlag, Berlin, Heidelberg.
- [4] Anderson, R.L. and Taflin, E., "Benjamin-Ono Equation-Recursivity of Linearization Maps - Lax Pairs", *Lett. in Math. Phys.* (in press).
- [5] Flato, M., Pinczon, G. and Simon, J., "Non Linear Representations of Lie Groups," *Ann. Scient. Éc. Norm. Sup.* 10, 405(1977).
- [6] Zakharov, V.E., and Faddeev, L.D., "Korteweg-de Vries equation, a completely integrable Hamiltonian system, *Funct. Anal. Appl.* 5, 280(1971).
- [7] Ablowitz, M. and Segur, H., "Solitons and the inverse scattering transform," SIAM, Philadelphia (1981).

- [8] Gardner, C.S., Greene, J.M., Kruskal, M.D., and Miura, R.M., "Method for solving the Korteweg-deVries equation, Phys. Rev. Lett. 19, 1095(1967).
- [9] Zakharov, V.E. and Shabat, P.B., "Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media, Sov. Phys. JETP 34, 62(1972).
- [10] Ablowitz, M.J., Kaup, D.J., Newell, A.C., and Segur H., "Method for solving the sine-Gordon equation," Phys. Rev. Lett. 30, 1262(1973).
- [11] Lax, P.D., "Integrals of nonlinear equations of evolution and solitary waves", Comm. Pure Appl. Math. 21, 467(1968).
- [12] Zakharov, V.E. and Mikhailov, A.V., "Relativistically invariant two-dimensional models of field theory which are integrable by means of the inverse scattering problem method", Sov. Phys. JETP 47, 1017(1978).
- [13] Wahlquist, H.D. and Estabrook, F.B., "Prolongation structures of nonlinear evolution equations", JMP 16, 1(1975).
- [14] Kac, V.G., Infinite Dimensional Lie Algebras, Birkhäuser, Boston (1984).
- [15] Pressley A. and Segal G., Loop groups and their representations. Oxford University Press, Oxford (in press).
- [16] Krichever, I.M., "Integration of nonlinear equation by methods of algebraic geometry". Funct. Anal Appl. 11, 12(1977).
- [17] Simon, J., "Survey on Non-Linear Representations of Lie Groups," in Proceedings AMS-SIAM Seminar on Applications of Group Theory in Phys. and Math. Phys., Univ. of Chicago (1980).
- [18] Taflin, E., "Dynamical Symmetries and Conservation Laws for the Korteweg-deVries Equation," Rep. Math. Phys. 20, 171(1984).
- [19] Taflin, E., "Analytic Linearization, Hamiltonian Formalism, and Infinite Sequences of Constants of Motion for the Burgers Equation", Phys. Rev. Lett. 47, 1425(1981).
- [20] Rosales, R., "Exact Solutions of Some Nonlinear Evolution Equations", Stud. Appl. Math. 59, 117(1978).
- [21] Fokas, A.S. and Ablowitz, M.J., "The Inverse Scattering Transform for the Benjamin-Ono Equation - A Pivot to Multidimensional Problems," Stud. Appl. Math. 68, 1(1983).
- [22] Hörmander, L., The Analysis of Linear Partial Differential Operators II, Springer-Verlag, Berlin, Heidelberg (1983).
- [23] Bock, T.L. and Kruskal, M.K., "A two-parameter Miura transformation of the Benjamin-Ono equation, Phys. Lett. 74A, 173(1979).

GROUP ANALYSIS OF THE PELLET FUSION PROCESS

V. J. Ervin, W. F. Ames
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332

and

E. Adams
Institut für Angewandte Mathematik
Universität Karlsruhe
75 Karlsruhe
Federal Republic of Germany

ABSTRACT. A gas dynamic model of the pellet fusion process having a time-invariant source term is studied. Under appropriate assumptions a nonlinear system of three partial differential equations results. A group analysis is performed on these equations and the family of one parameter transformation groups, which leave the equations invariant, is derived. Exploiting these invariants, for some particular values of the parameters involved in the equations, closed form solutions are found.

I. INTRODUCTION. In the pellet fusion process, a spherical pellet is fired into a containment chamber and then bombarded with pulses of laser energy. When hit by a laser pulse, a shock wave propagates through the pellet as the temperature rises. [A typical pellet configuration is sketched below.]

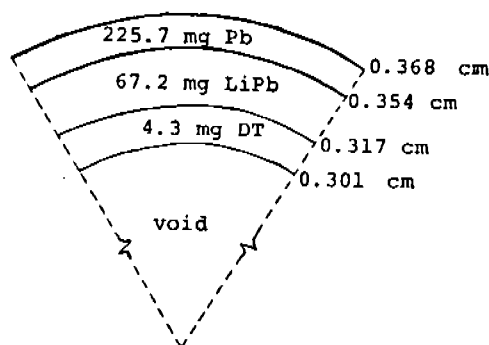


Fig. 1. A Typical Pellet Configuration.

The outermost layer of the pellet, consisting of lead, prevents the outward expansion of the inner layers; hence, as the Lithium-Lead layer heats up, it expands inward forcing the DT fuel toward the center.

The shock wave, after passing through to the inner layer, is reflected. On striking the outer surface of the pellet, another pulse arrives and constructively adds to the shock wave already in the pellet.

The process continues until the concentration of fuel and the temperature, at the center, is high enough for the fusion reaction to take place.

II. THE MODEL. By applying the principles of gas dynamics and exploiting the spherical symmetry of the problem, we obtain the following equations describing the process:

$$\frac{\partial \rho}{\partial t} + 2\rho u r^{-1} + \rho \frac{\partial u}{\partial r} + u \frac{\partial \rho}{\partial r} = 0 \quad (1)$$

(Conservation of Mass)

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial r} + R T \rho^{-1} \frac{\partial \rho}{\partial r} + R \frac{\partial T}{\partial r} = 0 \quad (2)$$

(Conservation of Momentum)

$$\frac{\partial T}{\partial t} + u \frac{\partial T}{\partial r} + 2(\gamma-1) T u r^{-1} + (\gamma-1) T \frac{\partial u}{\partial r} = r^n \quad (3)$$

(Conservation of Energy)

where $\rho(r,t)$ is the density inside the pellet ($\rho g/\mu m^3$)
 $u(r,t)$ is the velocity inside the pellet ($\mu m/ns$),
 $T(r,t)$ is the temperature inside the pellet (KeV),
 r is the radial variable (μm),
 t is the time (ns),
 R is the gas constant for lead ($\mu m^2/ns$ KeV),
 γ is the ratio of specific heat,
 r^n is the external energy input (KeV/ns).

The boundary/initial conditions are

$$\rho(r,0) = \rho_0(r), \quad u(r,0) = u_0(r), \quad T(r,0) = T_0(r)$$

and $u(0,t) = 0$.

In modeling the pellet fusion process, we have implicitly assumed

(a) the process is adiabatic

(b) the internal viscosity of the pellet is negligible.

III. THE FULL GROUP (Boisvert [3], Ervin [4]). The family of one parameter groups which leave equations (1)-(3) constant conformally invariant may be summarized as:

Theorem: The following three continuous one parameter transformations leave ()-() invariant.

$$Q_1: (r, t, \rho, u, T) \rightarrow (re^{\frac{1}{n+1}\epsilon}, te^{\frac{2-n}{3(n+1)}\epsilon}, \rho e^{-\frac{2n-1}{3(n+1)}\epsilon}, ue^{-\frac{1}{3}\epsilon}, Te^{\frac{2}{3}\epsilon})$$

$$Q_2: (r, t, \rho, u, T) \rightarrow (r, t+\epsilon, \rho, u, T)$$

$$Q_3: (r, t, \rho, u, T) \rightarrow (r, t, \rho e^\epsilon, u, T).$$

Proof: follows by direct substitution.

Note: Q_2 corresponds to the equations being translation invariant with respect to time. Q_1 and Q_3 indicates that the equations are invariant under a two parameter dilatation (or stretching) transformation.

IV. APPLICATION OF THE DILATATION GROUP (Ames [1], Bluman and Cole [2], Ovsiannikov [5]). Assume

$$\rho = e^{\alpha\bar{\rho}}, \quad u = e^{\beta\bar{u}}, \quad T = e^{\delta\bar{T}}, \quad r = e^{\epsilon\bar{r}}, \quad t = e^{\lambda\bar{t}}. \quad (4)$$

The invariants of the group are

$$\eta = \frac{t}{r^{\lambda/\epsilon}}; \quad f(\eta) = \frac{\rho}{r^{\alpha/\epsilon}}; \quad g(\eta) = \frac{u}{r^{\beta/\epsilon}}; \quad h(\eta) = \frac{T}{r^{\delta/\epsilon}}; \quad (5)$$

$$\epsilon \neq 0$$

The invariance of the P.D.E.'s (1)-(3) imply that

$$\frac{\beta}{\epsilon} = \frac{n+1}{3} \quad (6)$$

$$\frac{\delta}{\epsilon} = \frac{2}{3} (n+1) \quad (7)$$

$$\frac{\lambda}{\epsilon} = \frac{2-n}{3} \quad (8)$$

Following from the transformation described by (5) and subject to the condition (6)-(8), equations (1)-(3) become

$$f' + (2 + \frac{\beta}{\epsilon} + \frac{\alpha}{\epsilon})fg - \frac{\lambda}{\epsilon} \eta (fg)' = 0 \quad (9)$$

$$g' - \frac{\lambda}{\varepsilon} \eta (gg' + R \frac{f'}{f} h + Rh') + \frac{\beta}{\varepsilon} g^2 + (\frac{\alpha}{\varepsilon} + \frac{\delta}{\varepsilon}) Rh = 0 \quad (10)$$

$$h' - \frac{\lambda}{\varepsilon} \eta (gh' + (\gamma-1)g'h) + (\frac{\delta}{\varepsilon} + 2(\gamma-1) + \frac{\beta}{\varepsilon} (\gamma-1))gh = \beta_0, \quad (11)$$

where $f = f(\eta)$, $g = g(\eta)$, $h = h(\eta)$.

Now, let us investigate the boundary/initial conditions under the dilatation transformation

$$(i) \quad \rho(r,0) = \rho_0(r)$$

From (5) $\rho(r,t) = r^{\alpha/\varepsilon} f(\eta)$, so that $\rho(r,0) = r^{\alpha/\varepsilon} f(0)$.

Hence assuming that $\rho_0(r) = Ar^a$ gives $\alpha/\varepsilon = a$ and $f(0) = A$.

$$(ii) \quad u(r,0) = 0$$

Since $u(r,t) = r^{\beta/\varepsilon} g(\eta)$ it follows that $u(r,0) = r^{\beta/\varepsilon} g(0)$, whereupon $g(0) = 0$.

$$(iii) \quad T(r,0) = T_0(r)$$

Since $T(r,t) = r^{\delta/\varepsilon} h(\eta)$ then $T(r,0) = r^{2/3(n+1)} h(0)$.

Hence with $T_0(r) = Br^{2/3(n+1)}$ it follows $h(0) = B$.

$$(iv) \quad u(0,t) = 0$$

Recall $u(r,t) = r^{\beta/\varepsilon} g(\eta) = r^{\beta/\varepsilon} g(\frac{t}{r^{\lambda/\varepsilon}}) = r^{(n+1)/3} g(tr^{(n-2)/3})$, whereupon

$$\lim_{r \rightarrow 0} r^{(n+1)/3} g(tr^{(n-2)/3}) = 0$$

provided $n \geq 2$.

V. CONSTRUCTION OF EXACT SOLUTIONS. Consider the case when $n = 2$. Now $n = 2$ implies $\beta/\varepsilon = 1$; $\delta/\varepsilon = 2$; $\lambda/\varepsilon = 0$. Equations (9)-(11) then simplify to

$$f' + (3 + \frac{\alpha}{\varepsilon})fg = 0 \quad (12)$$

$$g' + g^2 + (2 + \frac{\alpha}{\varepsilon})Rh = 0 \quad (13)$$

$$h' + (3\gamma-1)gh = \beta_0 \quad (14)$$

with initial conditions $f(0) = A$, $g(0) = 0$, $h(0) = B$.

From (12) it follows that

$$f(\eta) = A \exp[-(3 + \frac{\alpha}{\varepsilon}) \int_0^\eta g(\xi) d\xi].$$

From (14)

$$h = - \frac{g' + g^2}{(2 + \frac{\alpha}{\epsilon})R} \quad (15)$$

whereupon

$$h' = - \frac{g'' + 2gg'}{(2 + \frac{\alpha}{\epsilon})R} \quad (16)$$

Equation (13) gives

$$g'(0) = -(2 + \frac{\alpha}{\epsilon})RB.$$

Substituting (15) and (16) into (14) yields

$$g'' + (3\gamma + 1)g'g' + (3\gamma - 1)g^3 = -\beta_0(2 + \frac{\alpha}{\epsilon})R \quad (17)$$

with initial condition

$$g(0) = 0, \quad g'(0) = -(2 + \frac{\alpha}{\epsilon})RB. \quad (18)$$

The change of variable

$$s' = \mu g s, \quad (19)$$

gives

$$s'' = \mu s(g' + \mu g^2) \quad (20)$$

and

$$s''' = \mu s(g'' + 3\mu g'g + \mu^2 g^3). \quad (21)$$

In particular, for $\gamma = 5/3$ (corresponding to a mono-atomic gas) and $\mu = 2$, equation (17) becomes

$$s''' + 2\beta_0(2 + \frac{\alpha}{\epsilon})Rs = 0 \quad (22)$$

with initial conditions

$$s(0) = 1, \quad s'(0) = 0, \quad s''(0) = -2(2 + \frac{\alpha}{\epsilon})RB. \quad (23)$$

The solution of (22) and (23) is

$$s(\eta) = c_1 e^{-F\eta} + e^{\frac{F}{2}\eta} (c_2 \cos \frac{\sqrt{3}}{2} F\eta + c_3 \sin \frac{\sqrt{3}}{2} F\eta),$$

where

$$F = (2\beta_0(2 + \frac{\alpha}{\epsilon})R)^{1/3}$$

and c_1, c_2, c_3 are given by

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 0 & \sqrt{3} & \sqrt{3} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -\frac{B}{\beta_0} F \end{bmatrix}.$$

From equation (19)

$$s(\eta) = \exp\left[2 \int_0^\eta g(\xi) d\xi\right]$$

and

$$\begin{aligned} g &= s'/2s \\ &= \frac{-F}{2} \frac{c_1 e^{-F\eta} + c_4 e^{\frac{F}{2}\eta} \sin\left(\frac{\sqrt{3}}{2} F\eta + \phi + \frac{\pi}{3}\right)}{c_1 e^{-F\eta} - c_4 e^{\frac{F}{2}\eta} \sin\left(\frac{\sqrt{3}}{2} F\eta + \phi\right)}, \end{aligned}$$

where

$$c_4 = \sqrt{c_2^2 + c_3^2} \quad \text{and} \quad \phi = \tan^{-1} \frac{c_2}{c_3}.$$

For $\gamma = 2/3$, we choose $\mu = 1$ in equation (19)-(21). Equation (17) then becomes

$$s''' + \beta_0 \left(2 + \frac{\alpha}{\epsilon}\right) R s = 0, \quad (24)$$

with initial conditions

$$s(0) = 1, \quad s'(0) = 0, \quad s''(0) = -\left(2 + \frac{\alpha}{\epsilon}\right) R B. \quad (25)$$

Letting

$$G = (\beta_0 (2 + \frac{\alpha}{\epsilon}) R)^{1/3}, \quad G \neq 0,$$

the solution for $s(\eta)$ is

$$s(\eta) = c_1 e^{-G\eta} + e^{\frac{G}{2}\eta} \left(c_2 \cos \frac{\sqrt{3}}{2} G\eta + c_3 \sin \frac{\sqrt{3}}{2} G\eta \right),$$

where the constants c_1, c_2, c_3 are given by

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 0 & \sqrt{3} & \sqrt{3} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -\frac{B}{\beta_0} G \end{bmatrix}.$$

Then $g(\eta)$ is given

$$g(\eta) = -G \frac{c_1 e^{-G\eta} + c_4 \frac{G}{e} \eta \sin(\frac{\sqrt{3}}{2} G\eta + \phi + \frac{\pi}{3})}{c_1 e^{-G\eta} - c_4 \frac{G}{e} \eta \sin(\frac{\sqrt{3}}{2} G\eta + \phi)}$$

where

$$c_4 = \sqrt{c_2^2 + c_3^2} \quad \text{and} \quad \phi = \tan^{-1} \frac{c_2}{c_3}.$$

In the case $\gamma = 1/3$, equation (17) becomes

$$g'' + 2g'g = -\beta_0 (2 + \frac{\alpha}{\epsilon}) R.$$

Integrating, we obtain the first order differential equation

$$g' + g^2 = -\beta_0 (2 + \frac{\alpha}{\epsilon}) R\eta + c. \quad (26)$$

Applying the initial conditions $g(0) = 0$ and $g'(0) = -(2 + \frac{\alpha}{\epsilon})RB$, we obtain $c = -(2 + \frac{\alpha}{\epsilon})RB$. Substituting this into equation (26) results in

$$g' + g^2 = -\beta_0 (2 + \frac{\alpha}{\epsilon}) R(\eta + \frac{B}{\beta_0}). \quad (27)$$

Upon applying equation (19) and (20) with $\mu = 1$, equation (27) becomes

$$s'' + \beta_0 (2 + \frac{\alpha}{\epsilon}) R(\eta + \frac{B}{\beta_0}) s = 0, \quad (28)$$

with initial conditions

$$s(0) = 1, \quad s'(0) = 0.$$

Making the change of variable $\xi = \eta + \frac{B}{\beta_0}$, equation (28) becomes

$$s'' + \beta_0 (2 + \frac{\alpha}{\epsilon}) R\xi s = 0, \quad (29)$$

with

$$s(\frac{B}{\beta_0}) = 1, \quad s'(\frac{B}{\beta_0}) = 0. \quad (30)$$

Equation (29) is a Bessel equation of order $1/3$, whose solution is

$$s(\xi) = \xi^{1/2} [c_1 J_{1/3}(\Omega(\xi)) + c_2 J_{-1/3}(\Omega(\xi))], \quad (31)$$

where

$$\Omega(\xi) = (\beta_0(2 + \frac{\alpha}{\epsilon})R)^{1/2} \frac{2}{3} \xi^{3/2}.$$

Differentiating yields

$$s'(\xi) = \frac{1}{2} \xi^{-1/2} [c_1 J_{1/3}(\Omega) + c_2 J_{-1/3}(\Omega)] \\ + (\beta_0(2 + \frac{\alpha}{\epsilon})R)^{1/2} \xi [c_1 J'_{1/3}(\Omega) + c_2 J'_{-1/3}(\Omega)],$$

that is,

$$s'(\xi) = \frac{1}{2} \xi^{-1/2} \{ [c_1 J_{1/3}(\Omega) + c_2 J_{-1/3}(\Omega)] \\ + 3\Omega [c_1 J'_{1/3}(\Omega) + c_2 J'_{-1/3}(\Omega)] \}.$$

For the application of initial conditions, it is convenient to set

$$\Omega_0 = \Omega(\frac{B}{\beta_0}) = (\beta_0(2 + \frac{\alpha}{\epsilon})R)^{1/2} \frac{2}{3} (\frac{B}{\beta_0})^{3/2}.$$

Then $s(B/\beta_0) = 1$ implies

$$c_1 J_{1/3}(\Omega_0) + c_2 J_{-1/3}(\Omega_0) = (\frac{B}{\beta_0})^{1/2} \quad (32)$$

and $s'(B/\beta_0) = 0$ gives

$$[c_1 J_{1/3}(\Omega_0) + c_2 J_{-1/3}(\Omega_0)] \\ + 3\Omega_0 [c_1 J'_{1/3}(\Omega_0) + c_2 J'_{-1/3}(\Omega_0)] = 0. \quad (33)$$

Using equation (32), (33) reduces to

$$c_1 J'_{1/3}(\Omega_0) + c_2 J'_{-1/3}(\Omega_0) = -\frac{1}{3\Omega_0} (\frac{\beta_0}{B})^{1/2}.$$

Hence c_1 and c_2 are given by

$$\begin{bmatrix} J_{1/3}(\Omega_0) & J_{-1/3}(\Omega_0) \\ J'_{1/3}(\Omega_0) & J'_{-1/3}(\Omega_0) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} (\frac{\beta_0}{B})^{1/2} \\ -\frac{1}{3\Omega_0} (\frac{\beta_0}{B})^{1/2} \end{bmatrix} \quad (34)$$

Observe that $\det \begin{bmatrix} J_{1/3}(\Omega_0) & J_{-1/3}(\Omega_0) \\ J'_{1/3}(\Omega_0) & J'_{-1/3}(\Omega_0) \end{bmatrix} =$

$$\text{Wronskian } (J_{1/3}(\Omega), J_{-1/3}(\Omega)) \Big|_{\Omega=\Omega_0}.$$

as the functions $J_{1/3}(\Omega)$ and $J_{-1/3}(\Omega)$ are linearly independent

solutions of (29) then the

$$\text{Wronskian } (J_{1/3}(\Omega), J_{-1/3}(\Omega)) \big|_{\Omega=\Omega_0} \neq 0.$$

Hence c_1 and c_2 are uniquely determined by (34). Since

$$\text{Wronskian } (J_{1/3}(\Omega), J_{-1/3}(\Omega)) \big|_{\Omega=\Omega_0} = \frac{-2 \sin \frac{\pi}{3}}{\pi \Omega_0},$$

then from (34) we have

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \frac{\pi \Omega_0}{\sqrt{3}} \begin{bmatrix} -J'_{1/3}(\Omega_0) & J_{-1/3}(\Omega_0) \\ J'_{1/3}(\Omega_0) & -J_{1/3}(\Omega_0) \end{bmatrix} \begin{bmatrix} (\frac{\beta_0}{B})^{1/2} \\ \frac{1}{3\Omega_0} (\frac{\beta_0}{B})^{1/2} \end{bmatrix}$$

Using the relationship

$$J_{r-1}(z) - J_{r+1}(z) = 2J'_r(z),$$

$$\begin{aligned} s'(\xi) = & \frac{1}{2} \xi^{-1/2} \{ c_1 [J_{1/3}(\Omega) + \frac{3}{2} \Omega J_{-2/3}(\Omega) - \frac{3}{2} \Omega J_{4/3}(\Omega)] \\ & + c_2 [J_{-1/3}(\Omega) + \frac{3}{2} \Omega J_{-4/3}(\Omega) - \frac{3}{2} \Omega J_{2/3}(\Omega)] \}, \end{aligned}$$

and therefore,

$$\begin{aligned} g(\eta) = & \frac{1}{2} \left(\eta + \frac{B}{\beta_0} \right)^{-1} \left\{ \frac{c_1 [J_{1/3}(\Omega) + \frac{3}{2} \Omega J_{-2/3}(\Omega) - \frac{3}{2} \Omega J_{4/3}(\Omega)]}{c_1 J_{1/3}(\Omega) + c_2 J_{-1/2}(\Omega)} \right. \\ & \left. + \frac{c_2 [J_{-1/3}(\Omega) + \frac{3}{2} \Omega J_{-4/3}(\Omega) - \frac{3}{2} \Omega J_{2/3}(\Omega)]}{c_1 J_{1/3}(\Omega) + c_2 J_{-1/3}(\Omega)} \right\}, \end{aligned}$$

where

$$\Omega = (\beta_0 (2 + \frac{\alpha}{\epsilon}) R)^{1/2} \frac{2}{3} \left(\eta + \frac{B}{\beta_0} \right)^{3/2}.$$

The "similarity" variable η , given by

$$\eta = \text{tr} \frac{n-2}{3}$$

becomes $\eta = t$ for $n = 2$.

Hence the solutions obtained above correspond to a 'separation of variables solution.'

$$\rho(r, t) = r^{\frac{\alpha}{\epsilon}} f(t)$$

$$u(r,t) = rg(t)$$

$$T(r,t) = r^2 h(t).$$

The substitution described by equation (19) enabled us to rewrite the non-linear differential equations (17) and (18) as linear equations of one higher order. The uniqueness of the solution of equations (22)-(23), (24)-(25), and (29)-(30) implies a unique solution of equations (17)-(18) for the cases $\gamma = 5/3$, $2/3$, and $1/3$.

VI. SINGULARITIES OF $g(\eta)$. For each of the cases discussed above the unknown function g was expressible as

$$g = \frac{s'}{\mu s}. \quad (35)$$

Likewise f and h can be written in terms of s ,

$$f = \frac{A}{s \left(3 + \frac{\alpha}{\epsilon}\right)}$$

and

$$h = \frac{(\mu-1)(s')^2 - \mu s''s}{\mu^2 \left(2 + \frac{\alpha}{\epsilon}\right) R s^2}.$$

In the case $\gamma = 1/3$, combining () and ()

$$h = \beta_0 \left(\eta + \frac{B}{\beta_0}\right).$$

From equation (34) we see that the solution to our system of equations (9)-(11) becomes singular whenever $s(\eta) = 0$. Physically, infinite values of $g(\eta)$ are not possible; however, the fact that theoretically it does become unbounded leads us to investigate the zeros of $s(\eta)$.

$$\text{For } \gamma = 5/3 \quad s(\eta) = c_1 e^{-F\eta} - c_4 e^{\frac{F}{2}\eta} \sin\left(\frac{\sqrt{3}}{2} F\eta + \phi\right)$$

$$F = (2\beta_0 \left(2 + \frac{\alpha}{\epsilon}\right) R)^{1/3}$$

$$\gamma = 2/3 \quad s(\eta) = c_1 e^{-G\eta} - c_4 e^{\frac{G}{2}\eta} \sin\left(\frac{\sqrt{3}}{2} G\eta + \phi\right);$$

$$G = (\beta_0 \left(2 + \frac{\alpha}{\epsilon}\right) R)^{1/3}$$

and for

$$\gamma = 1/3 \quad s(\eta) = (\eta + \frac{B}{\beta_0})^{1/2} [c_1 J_{1/3}(\Omega) + c_2 J_{-1/3}(\Omega)];$$

$$\Omega(\eta) = (\beta_0 (2 + \frac{\alpha}{\epsilon}) R)^{1/2} \frac{2}{3} (\eta + \frac{B}{\beta_0})^{3/2}.$$

In the first two cases it is obvious that $s(\eta)$ will have infinitely many zeros for $\eta > 0$. Actually as η becomes large these zeros will occur almost periodically with periods $4\pi/\sqrt{3} F$ and $4\pi/\sqrt{3} G$ respectively. For $\gamma = 1/3$ let us examine equation (31)

$$s''(\eta) + \beta_0 (2 + \frac{\alpha}{\epsilon}) R (\eta + \frac{B}{\beta_0}) s(\eta) = 0.$$

Unlike the previous two cases, in which the zeros occur almost periodically, the zeros of $s(\eta)$ (for $\gamma = 1/3$) occur "more frequently." To see this we apply the "Interlacing of Zeros Theorem" [cf. 6].

Observe that

$$w(\eta) = (\eta + \frac{B}{\beta_0})^{1/2} J_{1/3}((\beta_0 (2 + \frac{\alpha}{\epsilon}) R)^{1/2} \frac{2}{3} (\eta + \frac{B}{\beta_0})^{3/2})$$

is a solution to equation (28). For large ξ the zeros of $J_{1/3}(\xi)$ are separated by approximately π . As η increases the amount by which η must vary, such that

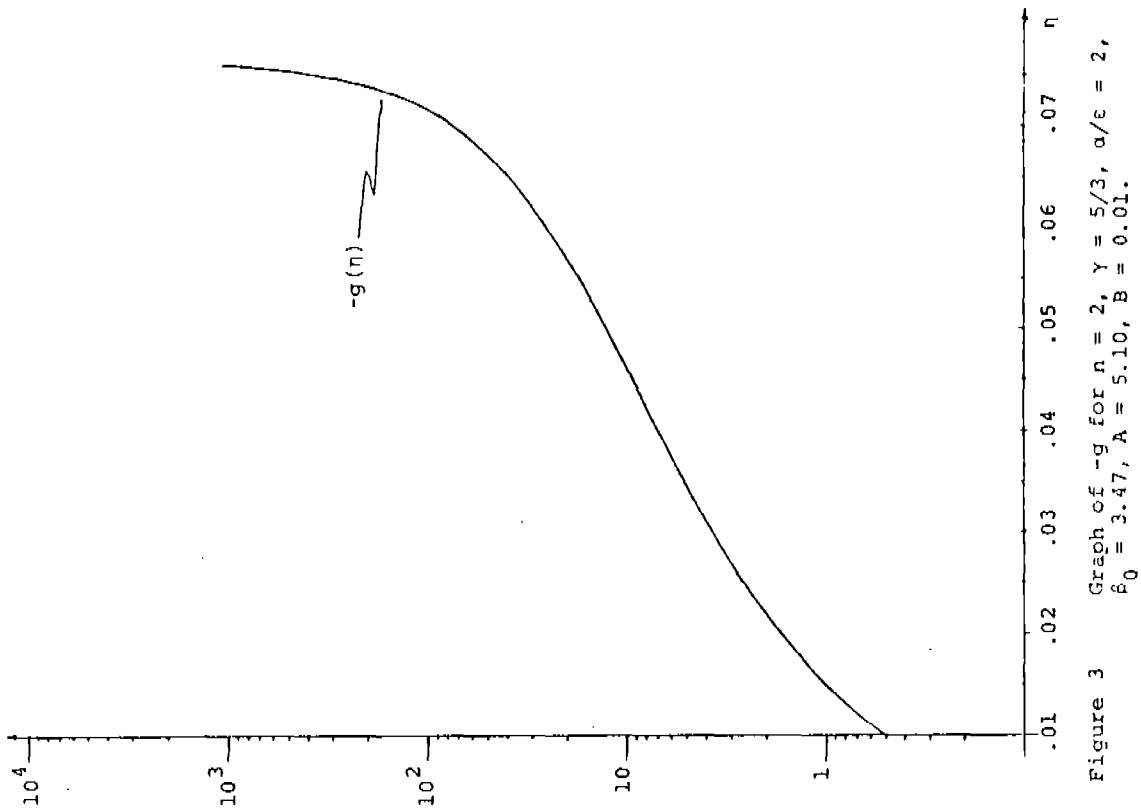
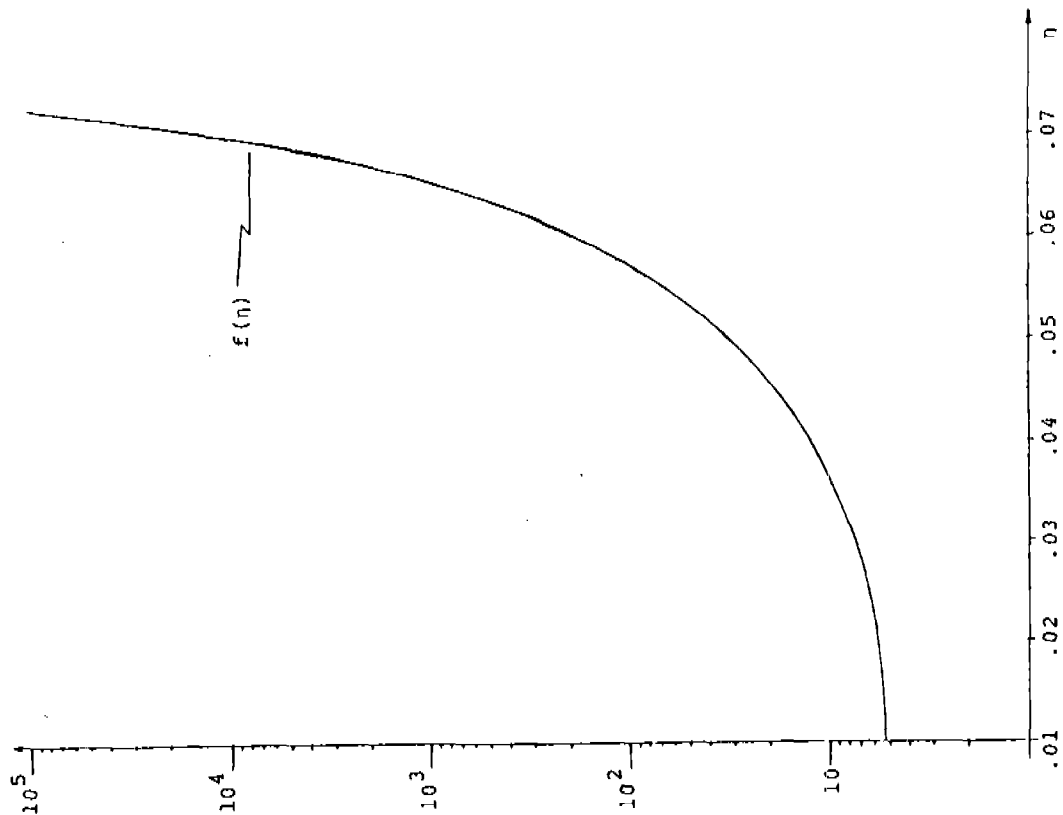
$$((\beta_0 (2 + \frac{\alpha}{\epsilon}) R)^{1/2} \frac{2}{3} (\eta + \frac{B}{\beta_0})^{3/2})$$

increases by π , decreases. Hence the zeros of $w(\eta)$ occur "more frequently" as η increases. Since the zeros of $s(\eta)$ and $w(\eta)$ are interlaced, it follows that the zeros of $s(\eta)$ must also occur "more frequently" as η increases.

References

1. Ames, W. F., Nonlinear Partial Differential Equations in Engineering, Vol. II, Academic Press, New York, 1972.
2. Bluman, G. W. and Cole, J. D., Similarity Methods for Differential Equations, Appl. Math. Sci., Vol. 13, Springer-Verlag, New York, 1974.
3. Boisvert, R. E., Group Analysis of the Navier-Stokes Equations, Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, 1982.
4. Ervin, V. J., Group Analysis of the Pellet Fusion Process, Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, 1984.

5. Ovsiannikov, L. V., Group Analysis of Differential Equations (in Russian), Moscow, 1978; English translation by W. F. Ames, Academic Press, 1982.
6. Birkhoff, G. and Rota, G., Ordinary Differential Equations, 3rd edition, John Wiley and Sons, New York, 1978, pp. 38-39.



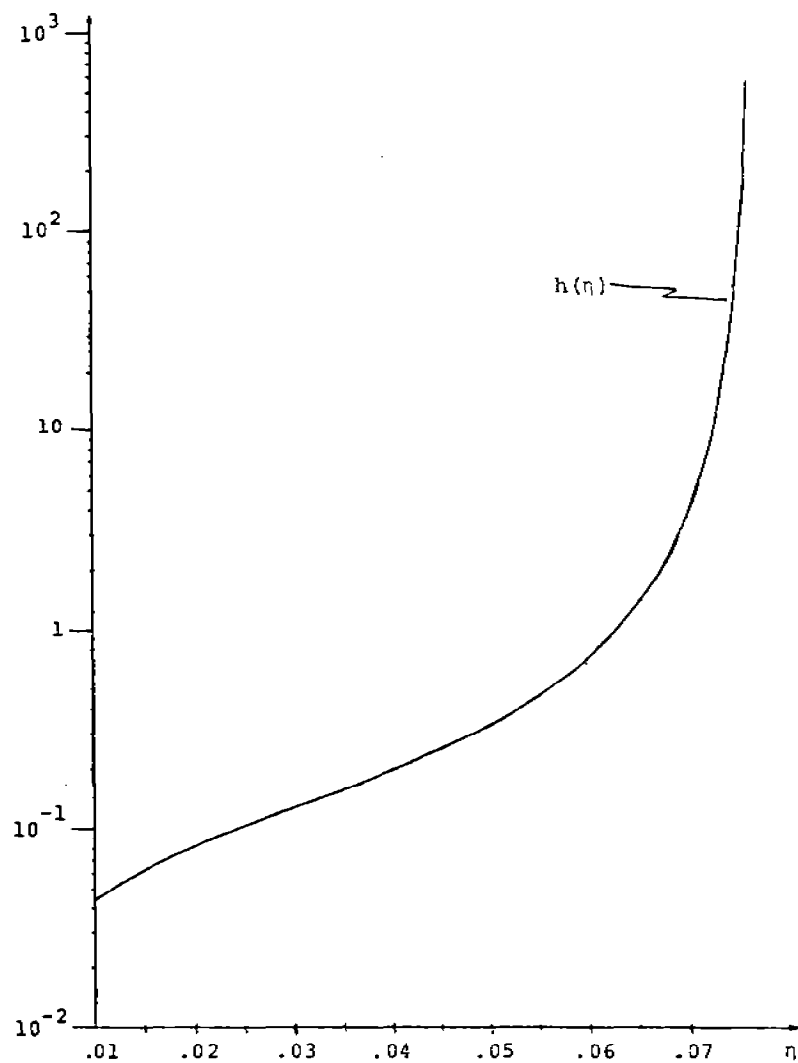


Figure 4. Graph of h for $n = 2$, $\gamma = 5/3$, $\alpha/c = 2$, $\beta_0 = 3.47$, $A = 5.10$, $B = 0.01$.

SATYA N. ATLURI

Center for the Advancement of Computational Mechanics
School of Civil Engineering
Georgia Institute of Technology

Abstract

(i) A synopsis of a recently developed methodology for computational small-deformation elasto-plastic analyses of cyclically loaded structures, based on a variant of an endochronic theory, is presented. The present plasticity model is shown to provide a unified basis for the earlier "multiple-yield-surface" or "nonlinear kinematic hardening" theories of Mroz, Krieg, Dafalias and Popov, Chaboche, and others. (ii) Issues related to constitutive modeling of finite deformation elasto-plasticity, and attendant numerical implementations, are discussed. Consistent rate forms of elastic-plastic evolution equations, which, in the limit, model the hyperelastic or small-deformation elastic-plastic behaviours appropriately, are presented.

1. An Endochronic Computational Approach to Cyclic Plasticity

Without much loss of generality, we treat only deviatoric plasticity. Let $\underline{\sigma}$ be the stress, $\underline{\sigma}'$ its deviator; $d\underline{\epsilon}$ the strain-rate, $d\underline{\epsilon}'$ the deviatoric strain; $d\underline{\epsilon}_m$ the mean strain, $d\underline{\epsilon}' = (d\underline{\epsilon}^e)' + d\underline{\epsilon}^p$; the plastic strain-change $d\underline{\epsilon}^p$ is purely deviatoric; and thus the differential of mean-strain, denoted by $d\underline{\epsilon}_m$, is purely elastic, i.e., $d\underline{\epsilon}_m \equiv d\underline{\epsilon}_m^e$. We consider the solid to be elastically isotropic. Using additive decomposition of $d\underline{\epsilon}$ into $d\underline{\epsilon}^e$ and $d\underline{\epsilon}^p$, we have:

$$d\underline{\epsilon}^p = d\underline{\epsilon}' - \frac{d\underline{\sigma}'}{2\mu} \quad (1.1a)$$

Following Valanis [1], we define endochronic (internal time) (but Newtonian time-like) parameters:¹

$$d\zeta = (d\underline{\epsilon}^p : d\underline{\epsilon}^p)^{1/2}; \quad d\zeta = \frac{d\zeta}{f(\zeta)}; \quad f(0) = 1, \quad d\zeta \geq 0 \quad (1.1b,c,d)$$

where $f(\zeta)$ is monotonically increasing.

As in Valanis [1], the stress in the elastic-plastic solid is represented through the integral:

¹ In what follows, $\underline{A} \cdot \underline{B} = A_{ij} B_{jk}$; and $\underline{A} : \underline{B} = A_{ij} B_{ij}$.

$$\underline{\sigma}' = 2\mu \int_0^z \rho(z-z') \frac{\partial \underline{\epsilon}^p}{\partial z'} dz' \quad (1.2)$$

where μ is the initial (elastic) shear modulus, and $\rho(z)$ is a material-specific kernel. Equation (1.2) thus appears to circumvent the need for a yield surface as well as for the flow rules of classical plasticity theory. Differentiation of (1.2) leads to:

$$d\underline{\sigma}' = 2\mu_p \left\{ d\underline{\epsilon}' + \frac{\underline{h}(z)}{\rho(0)f(z)} \left[\left(d\underline{\epsilon}' - \frac{d\underline{\sigma}'}{2\mu} \right) : \left(d\underline{\epsilon}' - \frac{d\underline{\sigma}'}{2\mu} \right) \right]^{1/2} \right\} \quad (1.3)$$

$$\rho(0) = \rho \text{ at } z = 0; \mu_p = \mu [1 + \rho(0)]^{-1}; \underline{h}(z) = \int_0^z \frac{\partial \rho}{\partial z} (z-z') \frac{\partial \underline{\epsilon}^p}{\partial z'} dz' \quad (1.4)$$

While the classical loading/unloading criteria (or criteria for elastic or plastic processes) are apparently bypassed in Eq. (1.3), there are, nevertheless, prices extracted for this seeming simplicity. Some of these counterbalancing difficulties of the above endochronic approach, as compared to a classical plasticity theory, are as follows: (i) The determination of stress history (and $d\underline{\sigma}$) for a given strain history (or $d\underline{\epsilon}$) at each material point becomes highly iterative in nature, as seen from (1.3); (ii) In a finite element/boundary-element/or other weak solution of the boundary value problem, the trial solution $d\underline{\epsilon}'$ is derived by differentiation of trial displacements $d\underline{u}$. To determine the trial stresses $d\underline{\sigma}'$ and yet retain a piecewise-linear-equation solution strategy, there is no recourse other than to approximate Eq. (1.3) as $d\underline{\sigma}' = 2\mu_p d\underline{\epsilon}'$. Thus, the stiffness matrix at any stage of loading is essentially the linear-elastic stiffness matrix; and the elastic-plastic solution method becomes the so-called "initial-strain" method; (iii) To model the uniaxial stress-strain curve of a material that does exhibit a sharp 'knee' near the elastic limit, the kernel $\rho(z)$ has to be weakly singular at $z = 0$. These drawbacks notwithstanding, Valanis and Fan [2] have recently presented a series of papers dealing with a direct computational implementation of an iterative, initial strain method based on Eq. (1.3) and using exponential functions for the kernel $\rho(z)$ in Eq. (1.4). Details of computational times for achieving convergence of plasticity iterations of the global finite element equations, or of the iterations for stress integration, are not readily available in [2].

Watanabe and Atluri [3-6] have recently presented alternate characterizations of cyclic-plasticity constitutive relations using the essential concepts of an endochronic theory, but with the following features: (i) The notion of a yield-surface, and the demarcation in the definitions of the elastic processes and plastic processes, are retained; (ii) The stress history (or $d\sigma$), for a given strain history (or $d\epsilon$), can be determined quite easily, as in a classical plasticity theory, by using a "generalized-midpoint-radial-return" algorithm; (iii) The finite-element formulation can be based on a "tangent-stiffness" approach, wherein the material constitutive law at each point can be chosen differently depending on whether an elastic process or a plastic process is postulated at each point during the current 'load' increment; and (iv) The present [3-6] endochronic approach provides a unified basis for other well-documented theories to model cyclic-plasticity, viz., the "multiple-yield-surface" theories of Mroz [7], Krieg [8], and Dafalias and Popov [9]; and the kinematic hardening theories of Prager, and Chaboche [12]. The starting point in the work of Watanabe and Atluri [3-4] is the representation of the kernel $\rho(z)$ in Eq. (1.2) in the form:

$$\rho(z) = \rho_0 \delta(z) + \rho_1(z) \quad (1.5)$$

where $\delta(z)$ is a Dirac function and $\rho_1(z)$ is a non-singular function. It turns out [3-4] that the term $\rho_0 \delta(z)$ in Eq. (1.5) leads to the notion of a yield surface; the function $f(\zeta)$ in (1.2b) leads to the notion of a yield-surface-expansion (isotropic hardening); and the function $\rho_1(z)$ in (1.5) leads to the notion of yield-surface translation (kinematic hardening). Use of (1.5) in (1.2) leads to:

$$\underline{\sigma}' = 2\mu \rho_0 \frac{d\underline{\epsilon}^p}{dz} + 2\mu \int_0^z \rho_1(z-z') \frac{d\underline{\epsilon}^p}{dz'} dz' \quad (1.6a)$$

$$\equiv \tau_y^0 \frac{d\underline{\epsilon}^p}{dz} + \underline{\sigma}'(z) \quad (1.6b)$$

wherein the definitions of τ_y^0 and $\underline{\sigma}'$ (the "back stress") are apparent. Eq. (1.6b) can be written as:

$$d\bar{\epsilon}^p = \frac{(\bar{g}' - \bar{\alpha}')}{\tau_y^0 f} \cdot d\zeta ; \quad d\zeta \geq 0 \quad (1.7)$$

Of course, Eq. (1.7) is entirely reminiscent of the classical flow-rule and normality relation for plastic strain-rate using a Mises' yield criterion. However, at this point, this similarity is purely formal.

From the very definition of $d\zeta$ as in (1.1b), it follows that, during plastic flow,

$$\frac{d\bar{\epsilon}^p}{d\zeta} : \frac{d\bar{\epsilon}^p}{d\zeta} = 1, \quad \text{i.e., } (\bar{g}' - \bar{\alpha}') : (\bar{g}' - \bar{\alpha}') = [\tau_y^0 f(\zeta)]^2 \quad (1.8a,b)$$

Equation (1.8b) clearly indicates that during plastic flow, the stress point, in the deviatoric stress space, remains on a Mises-cylinder of radius $\tau_y^0 f(\zeta)$, with the center of the surface at $\bar{\alpha}'$.

By differentiating (1.7) with respect to ζ , one obtains the following relation which holds during plastic flow:

$$\left(\frac{d^2 \bar{\epsilon}^p}{d\zeta^2} + \frac{d\bar{\epsilon}^p}{d\zeta} \frac{df}{d\zeta} \right) = \left(\frac{d\bar{g}'}{d\zeta} - \frac{d\bar{\alpha}'}{d\zeta} \right) \frac{1}{\tau_y} \quad (1.9)$$

From (1.1), and the definition of $\bar{\alpha}'$ as in (1.6), respectively, we see that:

$$\frac{d\bar{g}'}{d\zeta} = 2\mu \left(\frac{d\bar{\epsilon}'}{d\zeta} - \frac{d\bar{\epsilon}^p}{d\zeta} \right) \quad (1.10)$$

and

$$\frac{d\bar{\alpha}'}{d\zeta} = 2\mu \left[\rho_1(o) \frac{d\bar{\epsilon}^p}{d\zeta} + \frac{\bar{h}^*}{f} \right] \quad (1.11a)$$

$$\text{where } \bar{h}^* = \int_0^z \frac{d\rho_1}{dz} (z-z') \frac{\partial \bar{\epsilon}^p}{\partial z'} dz' \quad (1.11b)$$

Use of (1.10) and (1.11a) in (1.9) results in:

$$d\bar{\epsilon}' = \left[1 + \rho_1(o) + \tau_y^0 \frac{(df/d\zeta)}{2\mu} \right] d\bar{\epsilon}^p + \frac{\bar{h}^* d\zeta}{f} + \frac{\tau_y^0 f}{2\mu} \left(\frac{d^2 \bar{\epsilon}^p}{d\zeta^2} \right) d\zeta \quad (1.12)$$

Also, during plastic flow, it follows from (1.8a) and (1.7) that:

$$\frac{d^2 \underline{\epsilon}^P}{d\zeta^2} : \frac{d\underline{\epsilon}^P}{d\zeta} = 0 = \left(\frac{d^2 \underline{\epsilon}^P}{d\zeta^2} \right) : \left(\frac{\underline{\sigma}' - \underline{\alpha}'}{\tau_{yf}^0} \right) \quad (1.13)$$

Taking the trace of both sides of (1.12) with $[(\underline{\sigma}' - \underline{\alpha}')/(\tau_{yf}^0)]$ [or which is also equal to $(d\underline{\epsilon}^P/d\zeta)$] and using (1.13), one obtains:

$$d\underline{\epsilon}' : \frac{(\underline{\sigma}' - \underline{\alpha}')}{\tau_{yf}^0} = \left[1 + \rho_1(0) + \frac{\tau_{yf}^0 (df/d\zeta)}{2\mu} + \frac{h^* : (\underline{\sigma}' - \underline{\alpha}')}{\tau_{yf}^0} \right] d\zeta \equiv C d\zeta \quad (1.14)$$

wherein the definition of C is apparent. Eq. (1.14) can be rewritten as:

$$d\zeta = \frac{1}{C} \left[\frac{(d\underline{\epsilon}') : (\underline{\sigma}' - \underline{\alpha}')}{\tau_{yf}^0} \right] \equiv \frac{1}{C} d\underline{\epsilon}' : \underline{N} \quad (1.15)$$

where $\underline{N} = (\underline{\sigma}' - \underline{\alpha}')/\tau_{yf}^0$ is a unit "Normal".

Now, by definition, during a "Plastic Process", i.e. when $d\underline{\epsilon}^P \neq 0$, we have $d\zeta > 0$. Thus, (1.15) clearly indicates:

(A) Definition of a Plastic Process (P): $d\zeta > 0$

$$(P) \text{ if (i) } (\underline{\sigma}' - \underline{\alpha}') : (\underline{\sigma}' - \underline{\alpha}') = (\tau_{yf}^0)^2 ; \text{ and } d\underline{\epsilon}' : \underline{N} > 0 \quad (1.16)$$

Equation (1.15) also indicates that a "plastic process" is not possible if $\underline{N} : d\underline{\epsilon}' \leq 0$. In confirmity with this, we define an "elastic process" as follows:

(B) Definition of an Elastic Process (E): $d\zeta = 0$

$$(i) \text{ if } (\underline{\sigma}' - \underline{\alpha}') : (\underline{\sigma}' - \underline{\alpha}') < (\tau_{yf}^0)^2 ; \text{ or} \quad (1.17a)$$

$$(ii) \text{ if } (\underline{\sigma}' - \underline{\alpha}') : (\underline{\sigma}' - \underline{\alpha}') = (\tau_{yf}^0)^2 \text{ and } \underline{N} : d\underline{\epsilon}' \leq 0 \quad (1.17b)$$

It is interesting to observe that the (Elastic) and (Plastic) processes defined above, for the present endochronic theory, depend directly on whether $(\underline{N} : d\underline{\epsilon}) \gtrless 0$; while in the classical plasticity theory these processes depend, ab initio, on whether $(\underline{N} : d\underline{\sigma}) \gtrless 0$. In computational mechanics, the central problem of plasticity is to determine $d\underline{\sigma}$, given as $d\underline{\epsilon}$. In this context, the (E) and (P) criteria of (1.16) and (1.17) are more direct and more meaningful. Using (1.15) in (1.7), we obtain:

During (P):

$$d\epsilon^P = \frac{1}{C} \underline{N}(\underline{N}:d\epsilon') = \frac{1}{C} \underline{N}(\underline{N}:d\epsilon) \quad (1.18)$$

since \underline{N} is deviatoric. Recall that α' [see (1.6)] and \underline{h}^* [see (1.11b)]; and through them, the coefficient C [see (1.14)] depend on the kernel $\rho_1(z)$.

A convenient choice for the kernel $\rho_1(z)$ is:

$$\rho_1(z) = \sum_i \rho_{1i} \exp(-\beta_i z) \quad (1.19)$$

such that, from (1.6) it follows that:

$$\alpha' = \sum_i 2\mu_0 \int_0^z \{ \rho_{1i} \exp[-\beta_i(z-z')] \} \frac{d\epsilon^P}{dz'} dz' = \sum_i \alpha'^{(i)} \quad (1.20a)$$

$$\text{and } \underline{h}^* = \sum_i \left(\frac{-\beta_i}{2\mu_0} \right) \alpha'^{(i)} \quad (1.20b)$$

with $\alpha'^{(i)}$ being defined in an apparent fashion. From (1.20) it follows:

$$d\alpha' = 2\mu_0 \rho_1(0) d\epsilon^P - \left\{ \sum_i \frac{\beta_i \alpha'^{(i)}}{f} \right\} (d\epsilon^P : d\epsilon^P)^{1/2} \quad (1.21)$$

Thus, the evolution equation for α' is nonlinear in $d\epsilon^P$ and thus is similar to a nonlinear-kinematic-hardening relation [12]. It has been discussed in detail by Watanabe and Atluri [4] that the present theory, with the translation of the yield surface as in (1.20), and the expansion of the yield surface as specified by:

$$f = (1 + \gamma \zeta) \quad [\text{linear}], \quad (1.22a)$$

$$\text{or } f = a + (1 - a) \exp(-\psi \zeta) \quad [\text{saturated}], \quad (1.22b)$$

where γ and ψ are constants; and $\zeta = \int d\zeta$; includes the multiple-yield-surface theories of Mroz [7], Krieg [8], and Dafalias and Popov [9] as special cases.

Based on (1.18), the stress-strain relation in the present theory may be written as:

$$d\sigma' = 2\mu_0 [d\epsilon' - \gamma \frac{1}{C} \underline{N}(\underline{N}:d\epsilon')] \quad (1.23a)$$

$$(d\mathbf{g}:\mathbf{I}) = (2\mu + 3\lambda)(d\mathbf{e}:\mathbf{I})$$

where $\Gamma = 1$ in (P) and $\Gamma = 0$ in (E).

By assuming $\Gamma = 1$ or 0 appropriately, one may proceed to develop a tangent-stiffness finite-element method in the usual fashion. If the stress \mathbf{g}_n at state C_n , in an incremental solution, is known, the incremental stresses $\Delta\mathbf{g}$ corresponding to the trial-solutions $\Delta\mathbf{e}$ for incremental strains are determined in the usual fashion. We assume that \mathbf{g}_n is on the yield surface and further assume that the process had been plastic; i.e., $\|(\mathbf{g}'_n + 2\mu\Delta\mathbf{e}') - \mathbf{g}'_n\| > (\tau_y^0 f_n)$. Then, for any θ such that $0 < \theta < 1$, the algorithm for determining the actual stress-increment $\Delta\mathbf{g}$ in the plastic process proceeds as follows:

$$\mathbf{N}_\theta = \frac{(\mathbf{g}'_n + 2\mu\theta\Delta\mathbf{e}') - \mathbf{g}'_n}{\|(\mathbf{g}'_n + 2\mu\theta\Delta\mathbf{e}') - \mathbf{g}'_n\|} \quad (1.24a)$$

$$\Delta\mathbf{e}^P = \frac{1}{C_n} (\mathbf{N}_\theta : \Delta\mathbf{e}) \mathbf{N}_\theta ; \quad (1.24b)$$

$$\left\{ \text{where } C_n = \left[1 + \rho_1(0) + \frac{\tau_y^0 (d\mathbf{f}/d\zeta)}{2\mu_0} + \frac{\mathbf{h}^* : (\mathbf{g}' - \mathbf{g}')}{\tau_y^0 f} \right]_n \right\}$$

$$\Delta\mathbf{g}' = 2\mu \left[\Delta\mathbf{e}' - \frac{1}{C_n} (\mathbf{N}_\theta : \Delta\mathbf{e}) \mathbf{N}_\theta \right] \quad (1.24c)$$

$$\Delta\mathbf{g}:\mathbf{I} = (2\mu + 3\lambda)(\Delta\mathbf{e}:\mathbf{I}) \quad (1.24d)$$

$$\Delta\zeta = (\Delta\mathbf{e}^P : \Delta\mathbf{e}^P)^{1/2} ; \quad f_{n+1} = f_n + \Delta f \quad (1.24e)$$

$$\mathbf{g}'_{n+1} = \mathbf{g}'_n + \{ 2\mu \rho_1(0) \Delta\mathbf{e}^P - \left(\sum_i \frac{\beta_i \mathbf{a}'_n^{(i)}}{f_n} \right) \Delta\zeta \} \quad (1.24f)$$

Of course, several variants of the above algorithm, such as sub-incremental ones, are possible. The above tangent-stiffness finite-element, and generalized mid-point-radial-return-stress-integration, algorithm has been used by Watanabe and Atluri [5,6] to solve several problems of cyclic plasticity and non-proportional biaxial loading. It has been found that the present models capture the experimentally observed phenomena of cyclic hardening, cross-hardening, ratcheting, etc.

Because of the superior predictive capabilities of the present model and the fact that it is no more difficult to implement than the usual (classical) plasticity models, it may be a candidate for further exploitation in general purpose computational programs.

II. Towards a Consistent Finite Deformation Plasticity Theory and Attendant Computations:

In dealing with finite deformation problems, it has been (and to a large extent still is) customary in the computational mechanics literature to postulate constitutive theories in the form of linear relations between an objective rate (usually the Jaumann rate) of stress or an objective rate of an internal variable (such as the back stress in a kinematic hardening plasticity theory) on the one hand, and (the objective) velocity strain (or the symmetric part of velocity gradient) on the other. In this section, we reexamine such approaches and point out several alternative paths towards rational and consistent theories and attendant computational algorithms for finite deformation elasto-plasticity.

First we examine hyperelasticity, and then elastoplasticity, with the aim that a finite-deformation elastoplasticity theory should remain valid in the limits of hyperelasticity as well as of small-deformation elastoplasticity.

II.1 Hyperelasticity

We consider here an isotropic hyperelastic solid wherein: \underline{F} is the deformation gradient with polar decomposition $\underline{F} = \underline{R} \cdot \underline{U}_e = \underline{V}_e \cdot \underline{R}$, where \underline{U}_e and \underline{V}_e are the elastic stretches ($\underline{U}_e \equiv \underline{U}$ and $\underline{V}_e \equiv \underline{V}$ in this case); \underline{R} the rigid rotation; W_{0e} the elastic-strain-energy density per unit initial volume; \underline{g} the Kirchhoff stress ($= J \underline{\tau}$ where $J = \det \underline{F}$, and $\underline{\tau}$ the Cauchy stress). For observer transformations denoted by an orthogonal rotation \underline{Q} , it is well known that the various quantities transform as: $\underline{F} \rightarrow \underline{Q} \cdot \underline{F}$; $\underline{R} \rightarrow \underline{Q} \cdot \underline{R}$; $\underline{U}_e \rightarrow \underline{U}_e$; $\underline{V} \rightarrow \underline{Q} \cdot \underline{V}_e \cdot \underline{Q}^t$; and $\underline{g} \rightarrow \underline{Q} \cdot \underline{g} \cdot \underline{Q}^t$. Based on these observer-frame related transformations, one may write an objective constitutive relation for an isotropic hyperelastic solid in any number of alternative forms as follows:

$$\begin{aligned}\underline{\underline{g}} &= \underline{\underline{R}} \cdot \underline{\underline{J}} \underline{\underline{\Gamma}}(\underline{\underline{U}}_e) \cdot \underline{\underline{R}}^t = \underline{\underline{F}} \cdot \underline{\underline{S}}(\underline{\underline{U}}_e) \cdot \underline{\underline{F}}^t = \underline{\underline{F}}^{-t} \cdot \underline{\underline{C}}(\underline{\underline{U}}_e) \cdot \underline{\underline{F}}^{-1} = \underline{\underline{F}}^{-t} \cdot \underline{\underline{M}}(\underline{\underline{U}}_e) \cdot \underline{\underline{F}}^t = \\ &= \underline{\underline{F}} \cdot \underline{\underline{r}}(\underline{\underline{U}}_e) \cdot \underline{\underline{R}}^t = \underline{\underline{V}} \cdot \underline{\underline{T}}(\underline{\underline{V}}_e) = \underline{\underline{g}}(\underline{\underline{V}}_e)\end{aligned}\quad (2.1)$$

where $\underline{\underline{\Gamma}}$ is the "rotated" stress, $\underline{\underline{S}}$ is the 2nd Piola-Kirchhoff stress, $\underline{\underline{C}}$ the "convected" stress, $\underline{\underline{M}}$ is another "induced" convected stress, $\underline{\underline{r}}$ the "Biot-Lure-Jaumann" stress. The physical interpretations of these tensors, as well as that of $\underline{\underline{T}}$, are given in Atluri [13]. The tensors $\underline{\underline{I}}$, $\underline{\underline{S}}$, $\underline{\underline{C}}$, $\underline{\underline{M}}$, and $\underline{\underline{r}}$ are functions of $\underline{\underline{U}}_e$ alone and, hence, are observer invariant, while $\underline{\underline{T}}$ is a function of $\underline{\underline{V}}_e$ and thus transforms as $\underline{\underline{T}} \rightarrow \underline{\underline{Q}} \cdot \underline{\underline{T}} \cdot \underline{\underline{Q}}^t$. From (2.1) it is easy to note the following equalities:

$$\underline{\underline{J}} \underline{\underline{\Gamma}}(\underline{\underline{U}}_e) = \underline{\underline{U}}_e \cdot \underline{\underline{S}}(\underline{\underline{U}}_e) \cdot \underline{\underline{U}}_e = \underline{\underline{U}}_e^{-1} \cdot \underline{\underline{C}}(\underline{\underline{U}}_e) \cdot \underline{\underline{U}}_e^{-1} = \underline{\underline{U}}_e^{-1} \cdot \underline{\underline{M}}(\underline{\underline{U}}_e) \cdot \underline{\underline{U}}_e = \underline{\underline{U}}_e \cdot \underline{\underline{r}}(\underline{\underline{U}}_e) \quad (2.2)$$

As shown in detail in Atluri [13], for isotropic elastic solids, the following relations exist between the various stresses and the elastic-strain-energy density W_{oe} :

$$\underline{\underline{\Gamma}}(\underline{\underline{U}}_e) = \frac{\partial W_{oe}}{\partial \ln \underline{\underline{U}}_e} ; \quad \underline{\underline{S}}(\underline{\underline{U}}_e) = \frac{\partial W_{oe}}{\partial (\underline{\underline{U}}_e^2)} ; \quad \underline{\underline{C}}(\underline{\underline{U}}_e) = \frac{\partial W_{oe}}{\partial (\underline{\underline{U}}_e^{-2})} ; \quad \underline{\underline{r}}(\underline{\underline{U}}_e) = \frac{\partial W_{oe}}{\partial \underline{\underline{U}}_e} \quad (2.3)$$

$$\text{and} \quad \underline{\underline{T}}(\underline{\underline{V}}_e) = \frac{\partial W_{oe}}{\partial \underline{\underline{V}}_e} ; \quad \underline{\underline{g}}(\underline{\underline{V}}_e) = \frac{\partial W_{oe}}{\partial \ln \underline{\underline{V}}_e} \quad (2.4)$$

where $\ln(\underline{\underline{\cdot}})$ denotes the natural logarithm of $(\underline{\underline{\cdot}})$. For initially unstressed solids, we have the restrictions:

$$\underline{\underline{\Gamma}} = \underline{\underline{S}} = \underline{\underline{C}} = \underline{\underline{r}} = \underline{\underline{0}} \text{ when } \underline{\underline{U}}_e = \underline{\underline{I}} ; \text{ and } \underline{\underline{T}} = \underline{\underline{g}} = \underline{\underline{0}} \text{ when } \underline{\underline{V}}_e = \underline{\underline{I}} . \quad (2.5)$$

Based on (2.4) and (2.5), one may define a restricted class of "semi-linear" isotropic hyperelastic solid through the relations:

$$\underline{\underline{\Gamma}} = 2\mu \ln \underline{\underline{U}}_e + \lambda [(\ln \underline{\underline{U}}_e) : \underline{\underline{I}}] \underline{\underline{I}} \quad (2.5a)$$

$$\underline{\underline{S}} = 2\mu (\underline{\underline{U}}_e^2 - \underline{\underline{I}}) + \lambda [(\underline{\underline{U}}_e^2 - \underline{\underline{I}}) : \underline{\underline{I}}] \underline{\underline{I}} \quad (2.5b)$$

$$\underline{\underline{C}} = 2\mu (\underline{\underline{U}}_e^{-2} - \underline{\underline{I}}) + \lambda [(\underline{\underline{U}}_e^{-2} - \underline{\underline{I}}) : \underline{\underline{I}}] \underline{\underline{I}} \quad (2.5c)$$

$$\underline{\underline{r}} = 2\mu (\underline{\underline{U}}_e - \underline{\underline{I}}) + \lambda [(\underline{\underline{U}}_e - \underline{\underline{I}}) : \underline{\underline{I}}] \underline{\underline{I}} \quad (2.5d)$$

$$\underline{\underline{T}} = 2\mu (\underline{\underline{V}}_e - \underline{\underline{I}}) + \lambda [(\underline{\underline{V}}_e - \underline{\underline{I}}) : \underline{\underline{I}}] \underline{\underline{I}} \quad (2.5e)$$

$$\underline{\dot{g}} = 2\mu \ln \underline{V}_e + \lambda [(\ln \underline{V}_e) : \underline{I}] \underline{I} \quad (2.5f)$$

wherein it is possible that the coefficients μ and λ may have different numerical values in each of Eqs. (2.5a-f).

Let $\underline{L} = \dot{\underline{F}} \cdot \underline{F}^{-1}$ be the velocity gradient, $\underline{D} = (\underline{L})_s [= \frac{1}{2}(\underline{L} + \underline{L}^t)]$ be the velocity strain $[(\underline{L})_s$ and $(\underline{L})_a$ denote the symmetric and antisymmetric parts, respectively, of (\underline{L})]; $\underline{W} = (\underline{L})_a [= \frac{1}{2}(\underline{L} - \underline{L}^t)]$ be the spin; and $\underline{\Omega} = \dot{\underline{R}} \cdot \underline{R}^t$. Now, by differentiating Eqs (2.1) with respect to time, one easily obtains a set of objective rate relations:

$$\dot{\underline{g}} - \underline{\Omega} \cdot \underline{g} + \underline{g} \cdot \underline{\Omega} \equiv \dot{\underline{g}}_G = \underline{R} \cdot \dot{\underline{F}} \cdot \underline{R}^t \quad (2.6a)$$

$$\begin{aligned} \dot{\underline{g}} - \underline{W} \cdot \underline{g} + \underline{g} \cdot \underline{W} \equiv \dot{\underline{g}}_J &= \underline{R} \cdot \left\{ \underline{F} - \frac{1}{2}(\dot{\underline{U}}_e \cdot \underline{U}_e^{-1} - \underline{U}_e^{-1} \cdot \dot{\underline{U}}_e) \cdot \underline{F} \right. \\ &\quad \left. - \underline{F} \cdot \frac{1}{2}(\underline{U}_e^{-1} \cdot \dot{\underline{U}}_e - \dot{\underline{U}}_e \cdot \underline{U}_e^{-1}) \right\} \cdot \underline{R}^t \equiv \underline{R} \cdot \left\{ \underline{F} - \underline{\bar{W}} \cdot \underline{F} + \underline{F} \cdot \underline{\bar{W}} \right\} \cdot \underline{R}^t \end{aligned} \quad (2.6b)$$

$$\dot{\underline{g}} - \underline{L} \cdot \underline{g} - \underline{g} \cdot \underline{L}^t \equiv \dot{\underline{g}}_T = \underline{F} \cdot \dot{\underline{S}} \cdot \underline{F}^t; \quad \dot{\underline{g}} + \underline{L}^t \cdot \underline{g} + \underline{g} \cdot \underline{L} = \dot{\underline{g}}_R = \underline{F}^{-t} \cdot \dot{\underline{C}} \cdot \underline{F}^{-1} \quad (2.6c,d)$$

$$\dot{\underline{g}} + \underline{L}^t \cdot \underline{g} - \underline{g} \cdot \underline{L}^t = \underline{F}^{-t} \cdot \dot{\underline{M}} \cdot \underline{F}^t; \quad \dot{\underline{g}} - \underline{L} \cdot \underline{g} + \underline{g} \cdot \underline{\Omega} = \dot{\underline{g}}_A = \underline{F} \cdot \dot{\underline{r}} \cdot \underline{R}^t \quad (2.6e,f)$$

$$\dot{\underline{g}} = \dot{\underline{V}}_e \cdot \underline{T}(\underline{V}_e) + \underline{V}_e \cdot \dot{\underline{T}}(\underline{V}_e); \quad \dot{\underline{g}} = 2\mu \frac{d}{dt} (\ln \underline{V}_e) + \lambda \left[\frac{d}{dt} (\ln \underline{V}_e) : \underline{I} \right] \underline{I} \quad (2.7)$$

In (2.6b), $\underline{\bar{W}}$ is the observer-invariant form of spin:

$$\underline{\bar{W}} = \underline{R}^t \cdot \underline{W} \cdot \underline{R} - \underline{R}^t \cdot \dot{\underline{R}} = \underline{R}^t \cdot (\underline{W} - \underline{\Omega}) \cdot \underline{R} \quad (2.8)$$

Also, $\dot{\underline{g}}_G$ is the Green-Naghdi rate, $\dot{\underline{g}}_J$ the Jaumann rate, $\dot{\underline{g}}_T$ the Truesdell rate, $\dot{\underline{g}}_R$ the Rivlin rate, etc. However, using (2.2) and (2.1), all the rate equations (2.6a-f) may be written in a unified form as:

$$\underline{R}^t \cdot \dot{\underline{g}}_G \cdot \underline{R} \equiv \dot{\underline{f}} \equiv \frac{d}{dt} (\underline{U}_e \cdot \underline{S} \cdot \underline{U}_e) = \frac{d}{dt} (\underline{U}_e^{-1} \cdot \underline{C} \cdot \underline{U}_e^{-1}) \equiv \frac{d}{dt} (\underline{U}_e^{-1} \cdot \underline{M} \cdot \underline{U}_e) \equiv \frac{d}{dt} (\underline{U}_e \cdot \underline{r}) \quad (2.10)$$

where $\underline{R}^t \cdot \dot{\underline{g}}_G \cdot \underline{R}$ may be considered as the Green-Naghdi rate in the rotated frame. If the eigendirections of \underline{U}_e are represented by the matrix α and those of \underline{V}_e are represented by β , we have (see Atluri [13]):

$$\frac{d}{dt} (\ln \underline{U}_e) = \underline{\alpha}^t \cdot \frac{d}{dt} (\ln \underline{\lambda}) \cdot \underline{\alpha} + \dot{\underline{\alpha}}^t \cdot \ln \underline{\lambda} \cdot \underline{\alpha} + \underline{\alpha}^t \cdot \ln \underline{\lambda} \cdot \dot{\underline{\alpha}} \quad (2.11a)$$

$$\frac{d}{dt} (\ln \underline{V}_e) = \underline{\beta}^t \cdot \frac{d}{dt} (\ln \underline{\mu}) \cdot \underline{\beta} + \dot{\underline{\beta}}^t \cdot \ln \underline{\mu} \cdot \underline{\beta} + \underline{\beta}^t \cdot \ln \underline{\mu} \cdot \dot{\underline{\beta}} \quad (2.11b)$$

where $\underline{\lambda}$ and $\underline{\mu}$ are diagonal matrices consisting of the eigenvalues of \underline{U}_e and \underline{V}_e respectively, and $\frac{d}{dt} \ln \underline{\lambda} = \underline{\lambda}^{-1} \cdot \dot{\underline{\lambda}} = \dot{\underline{\lambda}} \cdot \underline{\lambda}^{-1}$. Thus when (2.11a) and (2.11b) are used in (2.5a) and (2.5f), respectively, the expressions for $\dot{\underline{\Gamma}}$ and $\dot{\underline{g}}$ become rather complicated. On the other hand, when the elastic stretches are small, i.e., $\underline{U}_e = \underline{I} + \underline{\epsilon}_U$ and $\underline{V}_e = \underline{I} + \underline{\epsilon}_V$ and where each of the components $(\epsilon_{ij})_U$ and $(\epsilon_{ij})_V$ is $\ll 1$, then

$$\ln \underline{U}_e = \underline{\epsilon}_U - \frac{\underline{\epsilon}_U^2}{2} + \frac{\underline{\epsilon}_U^3}{3} + \dots; \quad \ln \underline{V}_e = \underline{\epsilon}_V - \frac{\underline{\epsilon}_V^2}{2} + \frac{\underline{\epsilon}_V^3}{3} - \dots \quad (2.12a,b)$$

Thus, for moderately small elastic stretches, (2.5a), (2.5f), and (2.12) lead to:

$$\begin{aligned} \dot{\underline{\Gamma}} = & 2\mu [\dot{\underline{\epsilon}}_U - \frac{1}{2}(\dot{\underline{\epsilon}}_U \cdot \underline{\epsilon}_U + \underline{\epsilon}_U \cdot \dot{\underline{\epsilon}}_U) + \frac{1}{3}(\dot{\underline{\epsilon}}_U \cdot \underline{\epsilon}_U^2 + \underline{\epsilon}_U^2 \cdot \dot{\underline{\epsilon}}_U + \underline{\epsilon}_U \cdot \dot{\underline{\epsilon}}_U \cdot \underline{\epsilon}_U)] \\ & + \lambda [\{\dot{\underline{\epsilon}}_U - \frac{1}{2}(\dot{\underline{\epsilon}}_U \cdot \underline{\epsilon}_U + \underline{\epsilon}_U \cdot \dot{\underline{\epsilon}}_U) + \dots\} : \underline{I}] \underline{I} \end{aligned} \quad (2.13a)$$

and a similar relation for $\dot{\underline{g}}$ with $\dot{\underline{\epsilon}}_U$ being replaced by $\dot{\underline{\epsilon}}_V$. For arbitrary \underline{U}_e and \underline{V}_e , it appears more algebraically convenient to work with relations of the type (2.5b), (2.5d), and (2.5e). Considering, for instance, (2.5d) and (2.5e), one has from (2.10):

$$\underline{R}^t \cdot \dot{\underline{g}}_G \cdot \underline{R} = 2\mu (\dot{\underline{U}}_e \cdot \underline{U}_e + \underline{U}_e \cdot \dot{\underline{U}}_e) + \lambda [\dot{\underline{U}}_e (\underline{U}_e : \underline{I}) + \underline{U}_e (\dot{\underline{U}}_e : \underline{I})] - (2\mu + 3\lambda) \dot{\underline{U}}_e \quad (2.14)$$

or, alternatively,

$$\dot{\underline{g}} = 2\mu (\dot{\underline{V}}_e \cdot \underline{V}_e + \underline{V}_e \cdot \dot{\underline{V}}_e) + \lambda [\dot{\underline{V}}_e (\underline{V}_e : \underline{I}) + \underline{V}_e (\dot{\underline{V}}_e : \underline{I})] - (2\mu + 3\lambda) \dot{\underline{V}}_e \quad (2.15)$$

Now, consider the polar decomposition, $\underline{F} = \underline{R} \cdot \underline{U}_e = \underline{V}_e \cdot \underline{R}$. From this we have $\underline{L} = \dot{\underline{F}} \cdot \underline{F}^{-1} = \underline{\Omega} + \underline{R} \cdot \dot{\underline{U}}_e \cdot \underline{U}_e^{-1} \cdot \underline{R}^t = \dot{\underline{V}}_e \cdot \underline{V}_e^{-1} + \underline{V}_e \cdot \underline{\Omega} \cdot \underline{V}_e^{-1}$.

$$\text{Thus, } \underline{\Omega} = \underline{R} \cdot (\dot{\underline{U}}_e \cdot \underline{U}_e^{-1})_s \cdot \underline{R}^t = (\dot{\underline{V}}_e \cdot \underline{V}_e^{-1})_s + (\underline{V}_e \cdot \underline{\Omega} \cdot \underline{V}_e^{-1})_s \quad (2.16)$$

$$\underline{W} = \underline{\Omega} + \underline{R} \cdot (\dot{\underline{U}}_e \cdot \underline{U}_e^{-1})_a \cdot \underline{R}^t = (\dot{\underline{V}}_e \cdot \underline{V}_e^{-1})_a + (\underline{V}_e \cdot \underline{\Omega} \cdot \underline{V}_e^{-1})_a \quad (2.17)$$

By defining coordinate-invariant velocity-strain $\underline{\bar{D}}$, and spin $\underline{\bar{W}}$, such that:

$\bar{W} = R^t \cdot (W - \underline{\Omega}) \cdot R$ and $\bar{D} = R^t \cdot D \cdot R$, it is easy to derive from (2.16 and 2.17) that:

$$\dot{\underline{U}}_e = (\bar{D} + \bar{W}) \cdot \underline{U}_e = \underline{U}_e \cdot (\bar{D} - \bar{W}) \quad (2.18)$$

and $\dot{\underline{V}}_e = (\underline{D} + \underline{W}) \cdot \underline{V}_e - \underline{V}_e \cdot \underline{\Omega} = \underline{V}_e \cdot (\underline{D} - \underline{W}) + \underline{\Omega} \cdot \underline{V}_e \quad (2.19)$

Use of (2.18) and (2.19) in (2.14) and (2.15) results in consistent forms of 'objective' rate-constitutive relations for hyperelastic solids. The resultant rate equations clearly indicate the fallacy in the current practice wherein a finite deformation (even hyper-elastic) stress-strain law is simply expressed as a linear relation between an objective rate of stress (usually $\dot{\underline{\sigma}}_J$) and \underline{D} . For instance, it is currently common practice to write,

$$(\dot{\underline{\sigma}}_J) \text{ or } (\dot{\underline{\sigma}}_G) = 2\mu\underline{D} + \lambda(\underline{D}:\underline{I})\underline{I} \quad (2.20)$$

Equations (2.14 and 2.18) and (2.15 and 2.19) clearly show that (2.20) is not valid. In fact, as is well known, Eq. (2.20) with $\dot{\underline{\sigma}}_J$ leads to oscillatory shear stresses in a finite deformation simple shear test [14].

While the above hyperelastic rate relations have their own intrinsic reasons for being, our objective here is to use them as guides to postulate consistent rate-type finite-deformation elasto-plastic relations. This is pursued next.

II.2 Rate-Type Finite-Deformation Elasto-Plastic Stress-Strain Relations

The primary criteria we shall demand these relations to satisfy are: (1) The finite elasto-plastic rate relations should, in the limit, be valid in the cases of hyperelasticity as well as small-deformation elasto-plasticity; (2) They should obey the so-called Prager's [15] criterion. This criterion implies that when the objective stress rate vanishes [by virtue of the vanishing of the right-hand side of the equation with the said stress rate on the left-hand side], then the second invariant of the Kirchhoff stress (\underline{g}) should remain constant. Otherwise, in a J_2 -flow theory of plasticity, spurious plastic flow may result; (3) The classical plasticity concepts, such as Drucker's postulates regarding plastic work and plastic normality, should hold.

We will assume that the solid is elastically isotropic. The main theme adopted now is that: the stress in a finitely deformed elasto-plastic solid may be derived from an elastic-strain-energy density function W_{oe} , such that:-

$$\begin{aligned} \underline{T}(\underline{U}_e) &= \frac{\partial W_{oe}}{\partial \ln \underline{U}_e} ; \quad \underline{S}(\underline{U}_e) = \frac{\partial W_{oe}}{\partial (\underline{U}_e^2)} ; \quad \underline{C}(\underline{U}_e) = \frac{\partial W_{oe}}{\partial (\underline{U}_e^{-2})} ; \\ \underline{r}(\underline{U}_e) &= \frac{\partial W_{oe}}{\partial \underline{U}_e} ; \quad \underline{T}(\underline{V}_e) = \frac{\partial W_{oe}}{\partial \underline{V}_e} ; \quad \underline{q}(\underline{V}_e) = \frac{\partial W_{oe}}{\partial \ln \underline{V}_e} \end{aligned} \quad (2.21)$$

i.e., in a manner entirely analogous to Eqs. (2.3 and 2.4). Thus, in a finitely deformed elastic-plastic solid, the 'objective' relations for the stress rate are still given by equations of the type Eqs. (2.14 and 2.18) or Eqs. (2.15 and 2.19), provided now $\dot{\underline{U}}_e$ and $\dot{\underline{V}}_e$ are appropriately defined in terms of the kinematics of an elastic-plastic deformation. As to this, Lee [16] has originally suggested:

$$\underline{F} = \underline{F}_e \cdot \underline{F}_p \quad (2.22)$$

with subscripts e and p denoting 'elastic' and 'plastic', respectively. However, it is well known that \underline{F}_e in (2.22) is not unique, but can be determined to only within a rigid rotation. Fardshisheh and Onat [17] have suggested that this non-uniqueness can be remedied by requiring \underline{F}_e to be a pure, symmetric stretch tensor, \underline{V}_e . Thus,

$$\underline{F} = \underline{V}_e \cdot \underline{F}_p \quad (2.23)$$

By requiring \underline{V}_e to be similar to that in an otherwise hyperelastic solid, we see that under observer transformations, the following results hold:

$\underline{F} \rightarrow \underline{Q} \cdot \underline{F}$; $\underline{V}_e \rightarrow \underline{Q} \cdot \underline{V}_e \cdot \underline{Q}^t$; and $\underline{F}_p \rightarrow \underline{Q} \cdot \underline{F}_p$. Note that in (2.23), the rigid-rotation of a material element is not explicitly determined. With the apparent loss of some generality, however, we study here the decompositions:

$$\underline{F} = \underline{R} \cdot \underline{U}_e \cdot \underline{U}_p \equiv \underline{V}_e \cdot \underline{V}_p \cdot \underline{R} \quad (2.24)$$

where \underline{R} is the rigid rotation, \underline{U}_e and \underline{V}_e are 'elastic' stretches, and \underline{U}_p and \underline{V}_p are the 'plastic' stretch tensors. We require \underline{U}_e , \underline{U}_p , \underline{V}_e , and \underline{V}_p to be symmetric. Under observer transformations it then follows that: $\underline{U}_e \rightarrow \underline{U}_e$;

$\underline{U}_p \rightarrow \underline{U}_p$; $\underline{V}_e \rightarrow \underline{Q} \cdot \underline{V}_e \cdot \underline{Q}^t$; $\underline{V}_p \rightarrow \underline{Q} \cdot \underline{V}_p \cdot \underline{Q}^t$; and also $\underline{V}_e = \underline{R} \cdot \underline{U}_e \cdot \underline{R}^t$; and $\underline{V}_p = \underline{R} \cdot \underline{U}_p \cdot \underline{R}^t$.
Under the decomposition (2.23), we have:

$$\underline{L} = \dot{\underline{V}}_e \cdot \underline{V}_e^{-1} + \underline{V}_e \cdot \dot{\underline{F}}_p \cdot \underline{F}_p^{-1} \cdot \underline{V}_e^{-1}; \quad (2.25a)$$

$$\underline{D} = (\dot{\underline{V}}_e \cdot \underline{V}_e^{-1})_s + (\underline{V}_e \cdot \dot{\underline{F}}_p \cdot \underline{F}_p^{-1} \cdot \underline{V}_e^{-1})_s; \quad \underline{W} = (\dot{\underline{V}}_e \cdot \underline{V}_e^{-1})_a + (\underline{V}_e \cdot \dot{\underline{F}}_p \cdot \underline{F}_p^{-1} \cdot \underline{V}_e^{-1})_a \quad (2.25b,c)$$

From (2.25) it easily follows that:

$$\dot{\underline{V}}_e = (\underline{D} + \underline{W}) \cdot \underline{V}_e - \underline{V}_e \cdot (\dot{\underline{F}}_p \cdot \underline{F}_p^{-1}) = \underline{V}_e \cdot (\underline{D} - \underline{W}) - (\dot{\underline{F}}_p \cdot \underline{F}_p^{-1})^t \cdot \underline{V}_e \quad (2.26)$$

Now, the stress-power, or the rate of stress-work, in a finitely deformed elastic-plastic body (per unit initial volume) is given by:

$$\dot{\underline{W}}_0 = \underline{g} : \underline{D} \equiv \dot{\underline{W}}_{oe} + \dot{\underline{W}}_{op} \quad (2.27a)$$

Thus, if \underline{D} is decomposed into $\underline{D} = \underline{D}_e + \underline{D}_p$, we will have:

$$\dot{\underline{W}}_{oe} = \underline{g} : \underline{D}_e; \quad \dot{\underline{W}}_{op} = \underline{g} : \underline{D}_p \quad (2.27b,c)$$

The well-known postulates of Drucker, concerning $\dot{\underline{W}}_{op}$, may then be used to establish the normality of \underline{D}_p to the yield surface, expressed as a function of \underline{g} . This suggests the definitions:

$$\underline{D}_p = (\underline{V}_e \cdot \dot{\underline{F}}_p \cdot \underline{F}_p^{-1} \cdot \underline{V}_e^{-1})_s; \quad \underline{W}_p = (\underline{V}_e \cdot \dot{\underline{F}}_p \cdot \underline{F}_p^{-1} \cdot \underline{V}_e^{-1})_a \quad (2.28)$$

where \underline{D}_p and \underline{W}_p are the "plastic velocity-strain" and "plastic spin", respectively. Using (2.28) in (2.26), we have:

$$\dot{\underline{V}}_e = (\underline{D} + \underline{W}) \cdot \underline{V}_e - (\underline{D}_p + \underline{W}_p) \cdot \underline{V}_e \equiv \underline{V}_e \cdot (\underline{D} - \underline{W}) - \underline{V}_e \cdot (\underline{D}_p - \underline{W}_p) \quad (2.29)$$

$$= (\underline{D}_e + \underline{W}_e) \cdot \underline{V}_e \equiv \underline{V}_e \cdot (\underline{D}_e - \underline{W}_e) \quad (2.30)$$

The use of (2.29) in (2.15) will then result in an objective, elastic-plastic stress-strain rate relation for finite deformations. We now need evolution equations for \underline{D}_p and \underline{W}_p as defined in Eq. (2.28), which may be written down, using the general isotropic function representation theorems, such as due to Wang [18]. In this connection, the work of Loret [19] in representing directly the quantity $(\dot{\underline{F}}_p \cdot \underline{F}_p^{-1})$ is useful and noteworthy. From this single representation, Loret [19] determines both \underline{D}_p and \underline{W}_p .

However, there appears a slight inconsistency in the above approach. If the process is purely elastic (and thus the material behaviour approaches hyperelasticity) or when there is no plastic process as determined by a flow rule, it follows that $\underline{D}_p = 0$. This necessitates $(\dot{\underline{E}}_p \cdot \underline{F}_p^{-1})$ to be zero, in an elastic process. Thus, in an elastic process, (2.29) reduces to:

$$\dot{\underline{V}}_e = (\underline{D} + \underline{W}) \cdot \underline{V}_e \quad (2.31)$$

which does not agree with the hyperelastic relation, (2.19), i.e., $\dot{\underline{V}}_e = (\underline{D} + \underline{W}) \cdot \underline{V}_e - \underline{V}_e \cdot \underline{\Omega}$.

On the other hand, using the decomposition (2.24), we have:

$$\underline{L} = \dot{\underline{R}} \cdot \underline{R}^t + \underline{R} \cdot \dot{\underline{U}}_e \cdot \underline{U}_e^{-1} \cdot \underline{R}^t + \underline{R} \cdot \underline{U}_e \cdot \dot{\underline{U}}_p \cdot \underline{U}_p^{-1} \cdot \underline{U}_e^{-1} \cdot \underline{R}^t \quad (2.32)$$

$$\text{or } \underline{D} = \underline{R} \cdot (\dot{\underline{U}}_e \cdot \underline{U}_e^{-1})_s \cdot \underline{R}^t + \underline{R} \cdot (\underline{U}_e \cdot \dot{\underline{U}}_p \cdot \underline{U}_p^{-1} \cdot \underline{U}_e^{-1})_s \cdot \underline{R}^t \quad (2.33)$$

$$\text{and } \underline{W} = \dot{\underline{R}} \cdot \underline{R}^t + \underline{R} \cdot (\dot{\underline{U}}_e \cdot \underline{U}_e^{-1})_a \cdot \underline{R}^t + \underline{R} \cdot (\underline{U}_e \cdot \dot{\underline{U}}_p \cdot \underline{U}_p^{-1} \cdot \underline{U}_e^{-1})_a \cdot \underline{R}^t \quad (2.34)$$

Equations (2.33) and (2.34) may be rearranged as:

$$\underline{\bar{D}} = \underline{R}^t \cdot \underline{D} \cdot \underline{R} = (\dot{\underline{U}}_e \cdot \underline{U}_e^{-1})_s + (\underline{U}_e \cdot \dot{\underline{U}}_p \cdot \underline{U}_p^{-1} \cdot \underline{U}_e^{-1})_s \quad (2.35)$$

$$\underline{\bar{W}} = \underline{R}^t \cdot (\underline{W} - \underline{\Omega}) \cdot \underline{R} = (\dot{\underline{U}}_e \cdot \underline{U}_e^{-1})_a + (\underline{U}_e \cdot \dot{\underline{U}}_p \cdot \underline{U}_p^{-1} \cdot \underline{U}_e^{-1})_a \quad (2.36)$$

Now, the stress-power per unit initial volume may be written as:

$$\dot{\underline{W}}_0 = \underline{g} : \underline{D} = (J \underline{R} \cdot \underline{\Gamma} \cdot \underline{R}^t) : \underline{D} = J \underline{\Gamma} : (\underline{R}^t \cdot \underline{D} \cdot \underline{R}) = J \underline{\Gamma} : \underline{\bar{D}} \quad (2.37)$$

Thus, if $\underline{\bar{D}}$ is decomposed into $\underline{\bar{D}} = \underline{\bar{D}}_e + \underline{\bar{D}}_p$, we have $\dot{\underline{W}}_{0e} = J \underline{\Gamma} : \underline{\bar{D}}_e$; $\dot{\underline{W}}_{0p} = J \underline{\Gamma} : \underline{\bar{D}}_p$.

Now, the yield function may be expressed as:

$$f(\underline{\Gamma}, \underline{W}_{op}) = 0 \quad (2.38)$$

since the invariants of $J \underline{\Gamma}$ are the same as those of \underline{g} . The Drucker-normality would then apply to $\underline{\bar{D}}_p$ in the deviatoric space of $\underline{\Gamma}$.

Thus, we may define:

$$\underline{\bar{D}}_p = (\underline{U}_e \cdot \dot{\underline{U}}_p \cdot \underline{U}_p^{-1} \cdot \underline{U}_e^{-1})_s ; \quad \underline{\bar{W}}_p = (\underline{U}_e \cdot \dot{\underline{U}}_p \cdot \underline{U}_p^{-1} \cdot \underline{U}_e^{-1})_a \quad (2.39)$$

Using (2.39), (2.35), and (2.36), we have:

$$\dot{\underline{U}}_e = (\underline{\bar{D}} + \underline{\bar{W}}) \cdot \underline{U}_e - (\underline{\bar{D}}_p + \underline{\bar{W}}_p) \cdot \underline{U}_e \equiv \underline{U}_e \cdot (\underline{\bar{D}} - \underline{\bar{W}}) - \underline{U}_e \cdot (\underline{\bar{D}}_p - \underline{\bar{W}}_p) \quad (2.40)$$

$$\equiv (\underline{\bar{D}}_e + \underline{\bar{W}}_e) \cdot \underline{U}_e \equiv \underline{U}_e \cdot (\underline{\bar{D}}_e - \underline{\bar{W}}_e) \quad (2.41)$$

the use of (2.40) in (2.14) results in an objective elastic-plastic stress-strain relation that is consistant for finite deformations. We may use representation theorems [18] to write down a general expression for $(\dot{\underline{U}}_p \cdot \underline{U}_p^{-1})$ from which both $\underline{\bar{D}}_p$ and $\underline{\bar{W}}_p$ can be determined from (2.39), such that $\underline{\bar{D}}_p$ satisfies the normality condition. We may essentially follow Loret [19].

The development in (2.40), in conjunction with (2.14), will remain consistent in the limit of hyperelasticity. Thus, when $\underline{U}_p \equiv \underline{I}$, $\dot{\underline{U}}_p \equiv 0$, $\underline{\bar{D}}_p = 0 = \underline{\bar{W}}_p$, and thus $\dot{\underline{U}}_e \equiv (\underline{\bar{D}} + \underline{\bar{W}}) \cdot \underline{U}_e$, which agrees with the hyperelastic relation, (2.18).

Also, the present suggestion of an elastic-plastic constitutive relation based on (2.40) and (2.14) does satisfy the Prager condition [15]. Thus, when $\underline{R}^t \cdot \dot{\underline{g}}_G \cdot \underline{R} = 0$, it is clear that since the rate of stress in a rigidly rotating system is zero, the invariants of the Kirchhoff tensor remain constant. This is not the case when other stress-rate equations as in (2.6c,d,e,f) [i.e., rates in non-rigidly-spinning systems] are used and when the right-hand sides of (2.6c,d,e, and f) are consistently determined (through time differentiation) from (2.5b,c,d), respectively. If, on the other hand, one uses ad-hoc postulations, as is commonly done currently, with only a linear relation between an objective stress rate and \underline{D} , such as:

$$(\dot{\underline{g}}_T) \text{ or } (\dot{\underline{g}}_R) \equiv 2\nu[\underline{D} - \Lambda(\underline{N}:\underline{D})\underline{N}] + \lambda(\underline{D}:\underline{I})\underline{I} \quad (2.42)$$

then, $\dot{\underline{g}}_T$ or $\dot{\underline{g}}_R = 0$ implies that $\underline{D} \equiv 0$, and then $\underline{L} \equiv \underline{W}$, and hence $\dot{\underline{g}}_T$ and $\dot{\underline{g}}_R$ reduce to $\dot{\underline{g}}_J$. Thus, since $\dot{\underline{g}}_J$ is a spin-based rate, when $\dot{\underline{g}}_T$ or $\dot{\underline{g}}_R$ is zero in (2.42), the invariants of \underline{g} themselves remain constant, and hence Prager's condition is obeyed by the inconsistent relation (2.42).

To close the description of other plasticity features, such as kinematic hardening, we may introduce a Kirchhoff-stress-like internal variable $\underline{\alpha}$, which we may label as the finite deformation back-stress. If

$$\dot{\underline{\alpha}}_G = \dot{\underline{\alpha}} - \underline{\Omega} \cdot \underline{\alpha} + \underline{\alpha} \cdot \underline{\Omega} , \quad (2.43)$$

one may introduce an isotropic tensor function representation

$$\dot{\underline{\alpha}}_G = f(\underline{\alpha}, \underline{g}, \underline{D}) \quad (2.44)$$

Such general representations, and specific examples, were considered by Reed and Atluri [22], who attempted, successfully, to model the results of finite-torsion experiments. For instance, Reed and Atluri [21] demonstrated that an excellent agreement with the experimental results of Swift [22] for stresses and strains in a finite-torsion test, can be obtained from the simple model:

$$\dot{\underline{\alpha}}'_J = C_1 \underline{D}^P + C_2 \underline{\alpha}' \quad (2.45)$$

where $C_1 = \text{constant}$; $C_2 = C_2(\underline{\alpha}' : \underline{D}^P)$,

in conjunction with

$$\underline{\alpha}'_J = A_1 \underline{D} + A_2 \underline{\alpha}' \quad (2.46)$$

where $A_1 = \text{constant}$; $A_2 = A_2(\underline{\alpha}' : \underline{D})$.

Another "nonlinear-kinematic-hardening" model, which is a special case of the representation in (2.44), and a generalization of (1.21), may be generated by writing:

$$\dot{\underline{\alpha}}'_G = C \underline{D}_p - \beta_1 \underline{\alpha}' (\underline{D}_p : \underline{D}_p)^{1/2} \quad (2.47)$$

The primary reason for using such models as in (2.45) and (2.47) is that simple linear models of the type

$$\underline{\alpha}'_J = C_1 \underline{D}^P \quad (2.48)$$

can easily be shown to lead (see Atluri [14]) to oscillatory values for the components of $\underline{\alpha}$ in cases such as finite simple shear. If this linearity is insisted upon, non-oscillatory back stress $\underline{\alpha}$ may be generated by replacing the left-hand side of (2.48) with other objective rates such as $\dot{\underline{\alpha}}'_G$, $\dot{\underline{\alpha}}'_T$, or $\dot{\underline{\alpha}}'_C$, as shown by Atluri [14], or by creating other

stress rates such as based on the spin of material fibers aligned with the principal directions of \underline{g} , as done by Lee [23]. However, since modeling material behaviour under cyclic loading necessitates the use of nonlinear kinematic hardening models as discussed in Section I of this paper, it may be worthwhile to circumvent the arbitrariness surrounding the left-hand side of (2.48) [while retaining only the linear term in \underline{D}^P as in (2.48)] and concentrate instead on the right-hand side of (2.48) and consider more general representations involving \underline{D}^P , $\underline{\alpha}$, \underline{g} as in the right-hand sides of (2.44) and (2.47). This type of modeling the evolution of the back stress $\underline{\alpha}$, along with the use of Eqs. (2.40) and (2.14) for representing the evolution of Kirchhoff stress \underline{g} , may form a consistent basis for a finite-elasto-plastic theory. This remains to be verified through systematic computational modeling of available experimental data on finite plasticity.

References

1. Valanis, K.C.: Fundamental consequence of a new intrinsic time measure-plasticity as a limit of endochronic theory. Arch. Mech. 31 (1980) 171.
2. Valanis, K.C. and Fan, J.: Endochronic analysis of cyclic elasto-plastic strain fields in a notched plate. Jrl. Appl. Mech. 50. 4a (1983) 789.
3. Watanabe, O. and Atluri, S.N.: Constitutive modeling of cyclic plasticity and creep, using an internal time concept. Int. Jrl. Plast. (1985) (in press).
4. Watanabe, O. and Atluri, S.N.: Internal time, general internal variable, and multi-yield-surface theories of plasticity and creep: a unification of concepts. Int. Jrl. Plast. (1985) (in press).
5. Watanabe, O. and Atluri, S.N.: A new endochronic approach to computational elastoplasticity: example of a cyclically loaded cracked plate. Jrl. of Appl. Mech. (1985) (in press).

6. Watanabe, O. and Atluri, S.N.: Endochronic approach to computational plasticity: non-proportional multiaxial cyclic loading. (in preparation).
7. Mroz, Z.: An attempt to describe the behavior of metals under cyclic loads using a more general workhardening model. *Acta Mech.* 7 (1969) 199.
8. Krieg, R.D.: A practical two surface plasticity theory. *Jrl. Appl. Mech.* 42 (1975) 641.
9. Dafalias, Y.F. and Popov, E.P.: Plastic internal variable formalism of cyclic plasticity. *Jrl. Appl. Mech.* 43 (1976) 645.
12. Chaboche, J.L. and Rousselier, G.: On the plastic and viscoplastic constitutive equations, part I: rules developed with internal variable concept. In *Inelastic analysis and life prediction in elevated temperature design*. ASME-PVP 59 (1982) 33.
13. Atluri, S.N.: Alternate stress and conjugate strain measures, and mixed variational formulations involving rigid rotations, for computational analyses of finitely deformed solids, with application to plates and shells-I theory. *Comp. & Struct.* 18. 1 (1983) 93.
14. Atluri, S.N.: On constitutive relations at finite strain: hypo-elasticity and elasto-plasticity with isotropic or kinematic hardening. *Comp. Meth. Appl. Mech. & Engg.* 43 (1984) 137.
15. Prager, W.: An elementary discussion of definitions of stress rate. *Quart. Appl. Mech.* 18 (1961) 403.
16. Lee, E.H.: Elastic-plastic deformations at finite strains. *Jrl. Appl. Mech.* 37 (1969) 1.
17. Fardshisheh, F. and Onat, E.T.: Representation of elastoplastic behaviour by means of state variables. In *Problems of plasticity*. A. Sawczuk (ed.) (1974) 89.
18. Wang, C.C.: A new representation theorem for isotropic functions. Parts 1 and 2. *Arch. Rat. Mech. Analysis.* 36 (1970) 166.
19. Loret, B.: On the effects of plastic rotation in the finite deformation of anisotropic elastic-plastic solids. *Mech. of Mat.* 2 (1983) 287.
21. Reed, K.W. and Atluri, S.N.: "Constitutive modeling and computational implementation for finite strain plasticity. *Int. Jrl. Plast.* 1. 1 (1985) 63.
22. Swift, H.W.: Length changes in metals under torsional overstrain. *Engineering.* 163 (1947) 253.
23. Lee, E.H., Mallett, R.L., and Wertheimer, T.B.: Stress analysis for anisotropic hardening in finite-deformation plasticity. *Jrl. Appl. Mech.* 50 (1983) 554.

J. A. Nohel and M. Renardy
 Mathematics Research Center
 University of Wisconsin-Madison
 610 Walnut Street
 Madison, WI 53705

ABSTRACT. We discuss the motion of nonlinear-viscoelastic materials with fading memory in one space dimension. We formulate the mathematical problem, survey results for global existence of classical solution to the initial value problem if the data are sufficiently small, and discuss in detail the development of singularities in initially smooth solutions for large data.

1. INTRODUCTION AND DISCUSSION OF RESULTS. In this paper we discuss the motion of nonlinear viscoelastic materials with fading memory in one space dimension. We concentrate on viscoelastic solids and briefly remark on similar results for fluids. After formulating the mathematical problems, we survey results for global existence of classical solutions to the initial value problem, provided the initial data are sufficiently small. We then discuss in some detail the development of singularities in initially smooth solutions for large data.

We consider the longitudinal motion of a homogeneous one-dimensional body occupying an interval B in a reference configuration and having unit reference density. For simple materials, the stress σ at a material point x is a nonlinear functional of the entire history of the strain $\epsilon = u_x$ at the same point x (here u denotes the displacement). In this paper, we confine ourselves to the following model problem, which can be motivated as a natural generalization of Boltzmann's constitutive relation for linear viscoelasticity [1] (the derivation of similar results in a variety of other models will be discussed in a later paper)

$$\sigma(x, t) = \varphi(\epsilon(x, t)) + \int_{-\infty}^t a'(t-\tau)\psi(\epsilon(x, \tau))d\tau, \quad (1.1)$$

$$(x \in B, -\infty < t < \infty).$$

Here φ and ψ are given smooth functions $\mathbb{R} \rightarrow \mathbb{R}$ with

$$\varphi(0) = \psi(0) = 0, \varphi' > 0, \psi' > 0, \quad (1.2)$$

and for physical reasons the relaxation function $a : [0, \infty) \rightarrow \mathbb{R}$ is positive, decreasing, convex, and $a' \in L^1[0, \infty)$, where $'$ denotes the derivative. The conditions on a imply that the stress relaxes as time increases and that deformations which occurred in the distant past have less influence on the present stress than those which occurred more recently. Since only a' occurs in the equation, we may use the normalization $a(\infty) = 0$. In the rheological literature the relaxation function a is often taken to be a finite linear combination of decaying exponentials with positive coefficients obtained by a least square fit to experimental data.

When (1.1) is substituted into the balance of linear momentum, the following integrodifferential equation for the displacement u results

$$u_{tt} = \varphi(u_x)_x + a' * \psi(u_x)_x + f, \quad x \in B, t > 0. \quad (1.3)$$

Here $*$ denotes the convolution $(\alpha * \beta)(t) = \int_0^t \alpha(t-\tau)\beta(\tau)d\tau$, and f is the sum of an external body force and the history term $\int_{-\infty}^0 a'(t-\tau)\psi(u_x(x, \tau))_x d\tau$. An appropriate dynamical problem is to determine a smooth function $u : B \times (0, \infty) \rightarrow \mathbb{R}$ which satisfies (1.3) together with appropriate boundary conditions if B is bounded, and which at $t = 0$ satisfies prescribed initial conditions

$$u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x), \quad x \in B$$

for certain smooth functions u_0 and u_1 . To avoid technical complications we assume in the following that $f = 0$. We also restrict ourselves to the case of an unbounded body, $B = \mathbb{R}$ and we study the Cauchy problem

$$u_{tt} = \varphi(u_x)_x + a' * \psi(u_x)_x, \quad x \in \mathbb{R}, t > 0, \quad (1.4)$$

$$u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x), \quad x \in \mathbb{R}. \quad (1.5)$$

When $a' \equiv 0$ and φ satisfies (1.2), the body is purely elastic. In this case it is well known (see Lax [14], MacCamy and Mizel [15], Klainerman and Majda [13]), that in general the Cauchy problem (1.4), (1.5) does not have globally defined smooth solutions, no matter how smooth and small the initial data are chosen. The initially smooth solution u develops singularities (shock waves) in finite time.

If $a' \not\equiv 0$ and a satisfies the sign conditions above, the fading memory term in (1.4) introduces a weak dissipation mechanism. Significant insight into the strength of this mechanism was gained by the work of Coleman and Gurtin [2], who studied the growth and decay of acceleration waves in materials with memory. They showed that the amplitude $q(t)$ of an acceleration wave propagating into a homogeneously strained medium at rest satisfies a Bernoulli-Riccati ordinary differential equation. The coefficient of q^2 in this equation is proportional to a second order elastic modulus, which is given by φ'' in our model problem, and there is a linear damping term proportional to $a'(0)$. Thus the amplitude $q(t) = [u_{tt}]$ decays to zero as $t \rightarrow \infty$, provided $|q(0)|$ is sufficiently small. On the other hand, if $\varphi'' \neq 0$, then $q(t) \rightarrow \infty$ in finite time if $|q(0)|$ is large enough, and $q(0)$ is of a certain sign. They did not study existence of such solutions.

This suggests that, under appropriate assumptions on ϕ , ψ and a , the Cauchy problem (1.4), (1.5) should have unique, globally defined classical (C^2) solutions for sufficiently smooth and small initial data u_0, u_1 , while smooth solutions should develop singularities in finite time if the initial data are large in an appropriate sense. Such a global existence result for small data were recently established by Hrusa and Nohel [10] using delicate a priori estimates obtained by combining an energy method with properties of Volterra equations (even in the presence of a small body force). We refer to a recent survey [9] for earlier small data results on initial boundary value problems modelling the motion of finite viscoelastic bodies, and for technical simplifications of the analysis in the special cases $\phi \equiv \psi$ or $a(t) = e^{-t}$. For the global results the Cauchy problem is more difficult than the finite body problem because the Poincaré inequality is not available to estimate lower order derivatives from higher order derivatives.

The remainder of our discussion will deal with the formation of singularities in finite time from smooth solutions of the Cauchy problem (1.4), (1.5). For the special case $\phi \equiv \psi$, Markowich and Renardy [17] have obtained numerical evidence for the formation of shock fronts in finite time from large data, and Hattori [7] has shown that, if $\phi'' \not\equiv 0$ and if the body B is finite, then there are smooth initial data (which he does not characterize) for which the corresponding Dirichlet-initial value problem does not have a globally defined smooth solution. On the other hand, Hrusa [8] has shown that if ϕ is linear and only ψ is allowed to be nonlinear, then the Cauchy problem (1.4), (1.5) does have globally smooth solutions, even for large smooth data. Therefore, we shall restrict ourselves to the case when $\phi'' \neq 0$, at least over the range of the solution. The case when ϕ'' changes sign will require further refinements.

An essential ingredient in the analysis (which is also important for the global theory) is the following local existence result which is established by combining Banach's fixed point theorem on an appropriate function space with standard energy estimates and Sobolev's embedding theorem.

Proposition 1:

Assume that $\phi, \psi \in C^3(\mathbb{R})$ satisfy (1.2); assume $a, a', a'' \in L^1_{loc}[0, \infty)$, (*)
and there is a constant $\kappa > 0$ such that

$$\phi'(\xi) > \kappa, \xi \in \mathbb{R}.$$

Assume that $u_0 \in L^2_{loc}(\mathbb{R})$ and that $u'_0, u_1 \in H^2(\mathbb{R})$. Then the Cauchy problem
(1.4), (1.5) has a unique classical solution $u \in C^2(\mathbb{R} \times [0, T_0])$ defined on
a maximal interval $(0, T_0)$. If T_0 is finite, then

$$\sup_{\mathbb{R} \times [0, T_0)} [|u_{xx}(x, t)| + |u_{xt}(x, t)|] = \infty.$$

(*)

Here the square bracket means integrability up to 0. No sign condition on a are required, but $a'(0)$ finite is essential.

The proof of Proposition 1 is almost identical to that of Theorem 2.1 of [6], and we omit the details; only certain readily available energy estimates for lower order derivatives are needed. The characterization of the maximal interval of existence is established by combining the energy estimates obtained in [6] with a Gronwall inequality argument. We remark that the energy estimates used in the proof of Proposition 1 yield time-dependent bounds which cannot be used to obtain global estimates. These can only be constructed by taking advantage of the damping mechanism induced by the memory term under appropriate sign conditions on a and by assuming the initial data to be small (see [10] for details).

The assumptions concerning the kernel a in Proposition 1 imply that a' is absolutely continuous on $[0, \infty)$. Recently, Hrusa and Renardy [11] established a result similar to Proposition 1 (and proved a global existence result for small data for bounded bodies) under assumptions which permit a singularity in a' at $t = 0$ (e.g. $a'(t) \sim -t^{\alpha-1}$, $0 < \alpha < 1$ as $t \rightarrow 0^+$). Such singularities are relevant for certain popular models of viscoelastic materials.

Our main result on development of singularities for large enough data is

Theorem 1:

Let $\varphi, \psi \in C^3(\mathbb{R})$ satisfy (1.2) and assume $a, a', a'' \in L^1_{loc}[0, \infty)$. Assume that $\varphi''(0) \neq 0$. Then, for every $T_1 > 0$, we can choose initial data $u_0', u_1 \in C^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$ such that the maximal time interval of existence, given by Proposition 1, for the smooth solution of the Cauchy problem (1.4), (1.5) cannot exceed T_1 . More precisely, if $\sup_{x \in \mathbb{R}} |u_0'(x)|$ and $\sup_{x \in \mathbb{R}} |u_1(x)|$ are sufficiently small, while $u_0''(x)$ and $u_1'(x)$ assume sufficiently large values with appropriate signs, then there is some $t^* < T_1$ such that

$$\sup_{\mathbb{R} \times [0, t^*)} \{ |u_{xx}(x, t)| + |u_{xt}(x, t)| \} = \infty, \quad (1.6)$$

while

$$\sup_{\mathbb{R} \times [0, t^*)} \{ |u_x(x, t)| + |u_t(x, t)| \} < \infty \quad (1.7)$$

(and in fact, this latter quantity remains small).

In view of the analogy with hyperbolic conservation laws and the numerical evidence [17], it is to be expected that a blow-up as established by Theorem 1 will lead to the development of a shock front.

The method of the proof, sketched in section 2 is to show that the memory term is in fact of lower order than the elastic term $\varphi(u_x)_x$ and can be treated as a perturbation. While considerably more technical, the proof is a generalization of the approach of Lax [14] for showing the development of

singularities for the quasilinear wave equation

$$u_{tt} = \varphi(u_x)_x .$$

Theorem 1 was established independently by Dafermos [4] using an approach which is different from ours but similar in spirit. The result can also be established by modifying the results of F. John [12] and extending them to systems of quasilinear hyperbolic conservation laws which contain lower order source terms (F. John, private communications).

Similar results for first order model problems were derived by Malek-Madani and Nohel [16] and, using different methods, by Renardy [18] and Dafermos [3].

A particular case of the model equation studied in this paper leads to a model for shearing flows of viscoelastic fluids studied recently by Slemrod [20]. With $v(x,t)$ denoting the velocity of the fluid in simple shear, Slemrod studies the problem

$$\begin{aligned} v_t &= a^* \varphi(v_x)_x , \quad (x \in \mathbb{R}, t > 0) , \\ v(x,0) &= v_0(x) , \quad (x \in \mathbb{R}) . \end{aligned} \tag{1.8}$$

for the special case $a = e^{-t}$. Problem (1.8) leads to a Cauchy problem of the form (1.4), (1.5) after differentiation with respect to time. Then Theorem 1 can be used to get a blow-up result for this problem, like the result found by Slemrod for $a(t) = e^{-t}$. The global existence of solutions for small data follows from [5, Theorem 4.1]. Other popular models for viscoelastic fluids have been analyzed by the method used in this paper; the results will be published elsewhere.

2. Development of Shocks. In this section, we sketch the proof of Theorem 1 establishing the development of shocks from initially smooth solutions of the Cauchy problem (1.4), (1.5) in finite time. For simplicity, most of the analysis will be carried out for the special case $a(t) = e^{-t}$; the proof for more general relaxation functions as well as for a more general class of model equations will be carried out in a forthcoming paper.

We begin by transforming (1.4) to an equivalent system. We let $w = u_x$, $v = u_t$, and write the constitutive assumption (1.1) in the form

$$\sigma = \varphi(w) - z , \quad z = -a^* \psi(w) .$$

Since we have assumed $\varphi' > 0$, the first of these equations can be solved for w ,

$$w = \varphi^{-1}(\sigma + z) =: g(\sigma, z) ,$$

and g is a smooth function of $\sigma \in \mathbb{R}$, $z \in \mathbb{R}$. As long as the solution remains smooth, the Cauchy problem (1.4), (1.5) is equivalent to the first order system

$$v_t = \sigma_x ,$$

$$\sigma_t = c^2(\sigma, z) v_x + a'(0) \psi(g(\sigma, z)) + a'' \psi(g(\sigma, z)) , \quad (2.1)$$

$$z_t = -a'(0) \psi(g(\sigma, z)) - a'' \psi(g(\sigma, z)) .$$

The initial conditions become

$$v(x, 0) = u_1(x), \quad \sigma(x, 0) = \varphi(u_0'(x)), \quad z(x, 0) = 0 . \quad (2.2)$$

By c we have denoted the wave speed

$$c(\sigma, z) := [\varphi'(g(\sigma, z))]^{1/2} ;$$

c is a smooth function of σ and z . The system (2.1) is hyperbolic, and its eigenvalues are $+c$, $-c$ and 0 . Under the assumptions of Proposition 1, a C^1 -solution exists on some maximal interval $R \times [0, T_0)$. If T_0 is finite, then v , σ , z or one of their first derivatives must become infinite as $t \rightarrow T_0$. It is immediate from equation (2.1) that σ_t , σ_x , z_t and z_x will remain bounded as long as v , σ , z , v_t and v_x are bounded.

To proceed further, we extend the classical approach of Lax [14] for first order hyperbolic 2×2 -systems. We define "approximate" Riemann invariants by those quantities which would be the classical Riemann invariants if z in the first two equations of (2.1) were treated as a parameter. These quantities are given by

$$\begin{aligned} r &= r(v, \sigma, z) = v + \Phi(\sigma, z) , \\ s &= s(v, \sigma, z) = v - \Phi(\sigma, z) , \\ \Phi(\sigma, z) &= \int_{\sigma_0}^{\sigma} \frac{d\zeta}{c(\zeta, z)} ; \end{aligned} \quad (2.3)$$

without loss of generality we may take $\sigma_0 = 0$. Since $\Phi_{\sigma}(\sigma, z) = \frac{1}{c(\sigma, z)} > 0$, this correspondence is smoothly invertible, and we have

$$v = \frac{r+s}{2} , \quad \Phi(\sigma, z) = \frac{r-s}{2} .$$

In the following, we assume $a(t) = e^{-t}$. Then (2.1) takes the simple form

$$\begin{aligned} v_t &= \sigma_x , \\ \sigma_t &= c^2(\sigma, z) v_x - \psi(g(\sigma, z)) + z , \\ z_t &= \psi(g(\sigma, z)) - z . \end{aligned} \quad (2.4)$$

We now differentiate r and s along the c and $-c$ characteristics, respectively, and z along the zero characteristic (i.e. we form $r_t - cr_x$, $s_t + cs_x$ and z_t). This leads to the following first order

hyperbolic system equivalent to (2.4), (2.2)

$$\begin{aligned}r_t - Ar_x &= -Bz_x + CD, \\s_t + As_x &= -Bz_x - CD, \\z_t &= D,\end{aligned}\tag{2.5}$$

with the initial data

$$\begin{aligned}r(x,0) &= u_1(x) + \Phi(\varphi(u_0'(x)),0), \\s(x,0) &= u_1(x) - \Phi(\varphi(u_0'(x)),0), \\z(x,0) &= 0;\end{aligned}\tag{2.6}$$

$$\begin{aligned}A &= A(r,s,z) := c(\sigma(r,s,z),z) > 0, \\B &= B(r,s,z) := c(\sigma(r,s,z),z)\Phi_z(\sigma(r,s,z),z), \\C &= C(r,s,z) := \Phi_z(\sigma(r,s,z),z) - \frac{1}{c(\sigma(r,s,z),z)}, \\D &= D(r,s,z) := \psi(g(\sigma(r,s,z),z)) - z.\end{aligned}\tag{2.7}$$

To establish the development of shocks in finite time, we study the evolution along characteristics of the quantities

$$\begin{aligned}\rho &:= v_x + \frac{\sigma_x}{c(\sigma,z)}, \\\tau &:= v_x - \frac{\sigma_x}{c(\sigma,z)},\end{aligned}\tag{2.8}$$

and z_x . Note that if z were a constant parameter, then ρ and τ would be the x -derivatives of r and s . We have $v_x = \frac{1}{2}(\rho + \tau)$, $\sigma_x = \frac{1}{2}c(\rho - \tau)$, and

$$(c^2)_\sigma(\sigma,z) = 2cc_\sigma = \frac{\Phi''(g(\sigma,z))}{\Phi'(g(\sigma,z))}.$$

A tedious but straightforward calculation using the relations (obtained by differentiating (2.4))

$$\begin{aligned}
v_{tx} &= \sigma_{xx} \quad , \\
\sigma_{tx} &= c^2(\sigma, z) v_{xx} + (c^2)_{\sigma}(\sigma, z) \sigma_x v_x \\
&\quad + (c^2)_z(\sigma, z) z_x v_x - D_x \quad , \\
z_{tx} &= D_x \quad ,
\end{aligned} \tag{2.9}$$

yields the system

$$\begin{aligned}
\rho_t - c\rho_x &= \frac{(c^2)_{\sigma}}{4} \rho(\rho-\tau) + O(|\rho||z_x| + |\tau||z_x| \\
&\quad + |\rho| + |\tau| + |z_x|) \quad , \\
\tau_t + c\tau_x &= -\frac{(c^2)_{\sigma}}{4} \tau(\rho-\tau) + O(|\rho||z_x| + |\tau||z_x| \\
&\quad + |\rho| + |\tau| + |z_x|) \quad , \\
z_{xt} &= O(|\rho| + |\tau| + |z_x|) \quad .
\end{aligned} \tag{2.10}$$

subject to the initial data

$$\begin{aligned}
\rho(x, 0) &= u_1'(x) + \varphi'(u_0'(x))^{1/2} u_0''(x) \quad , \\
\tau(x, 0) &= u_1'(x) - \varphi'(u_0'(x))^{1/2} u_0''(x) \quad , \\
z_x(x, 0) &= 0 \quad .
\end{aligned} \tag{2.11}$$

The cross product terms $\rho\tau$ in (2.10) are eliminated if one considers the characteristic derivatives of $c(\sigma, z)^{1/2}\rho$ and $c(\sigma, z)^{1/2}\tau$ (see Lax [14] and Slemrod [19]). We find

$$\begin{aligned}
\left(\frac{\partial}{\partial t} - c \frac{\partial}{\partial x}\right)(c^{1/2}\rho) &= \gamma(c^{1/2}\rho)^2 + O(|\rho||z_x| + |\tau||z_x| \\
&\quad + |\rho| + |\tau| + |z_x|) \quad , \\
\left(\frac{\partial}{\partial t} + c \frac{\partial}{\partial x}\right)(c^{1/2}\tau) &= \gamma(c^{1/2}\tau)^2 + O(|\rho||z_x| + |\tau||z_x| \\
&\quad + |\rho| + |\tau| + |z_x|) \quad , \\
z_{xt} &= O(|\rho| + |\tau| + |z_x|) \quad .
\end{aligned} \tag{2.12}$$

Here the coefficient function γ is given by

$$\gamma = \gamma(\sigma, z) = \frac{1}{4} \frac{\varphi''(g(\sigma, z))}{\varphi'(g(\sigma, z))^{5/4}}.$$

For definiteness, let us assume $\varphi''(0) > 0$ (the discussion for $\varphi''(0) < 0$ is analogous). We take initial data with the following properties: u_0^0 and u_1 (and hence $\rho(x, 0)$, $\tau(x, 0)$ as well as $z(x, 0) \equiv 0$) are uniformly small, and $\rho(x, 0)$, $\tau(x, 0)$ are such that at least one of them has a large positive maximum (by choosing u_0^0 or u_1 or both sufficiently large). At the same time, the maxima of $-\rho$ and $-\tau$ should not be too large.

As long as (r, s, z) remains within a given neighborhood U of 0 , we have upper and lower bounds for the coefficients occurring in (2.12), in particular, we have a positive lower bound γ_0 for γ . We shall see later that (r, s, z) will remain in U up to the time of blow-up if they are small enough initially and if we make the maximum of $\rho(x, 0)$ or $\tau(x, 0)$ large enough.

For every $t > 0$, we now set

$$h(t) = \max_x [\max_x \rho(x, t), \max_x \tau(x, t)].$$

From (2.12), we find that, as long as $(r, s, z) \in U$, while $h(t)$ is large and $\max_x |z_x| \ll h(t)$, we have, for some positive constants γ_0 and κ

$$\left(\frac{d}{dt}\right)_+ h(t) > \gamma_0 (h(t))^2, \text{ and } \max_x |z_{xt}| \leq \kappa h(t) \ll (h(t))^2.$$

Initially, we have $|z_x| = 0$ and it follows from these inequalities that it will remain small compared to $h(t)$. We also find that $h(t)$ becomes infinite in finite time. Since there is also some constant γ_1 such that $\left(\frac{d}{dt}\right)_+ h(t) \leq \gamma_1 (h(t))^2$, it can be shown that, with t^* denoting the blow-up time of h , we have $\frac{c_1}{t^* - t} \leq h(t) \leq \frac{c_2}{t^* - t}$ for some constants c_1 and c_2 .

The third equation of (2.12) then implies that $|z_x|$ grows at most logarithmically as $t \rightarrow t^*$. Since $\log(t^* - t)$ is integrable, equations (2.5) imply that r , s , and z remain bounded and in fact small if their initial data are small, and t^* is small (which is the case if $h(0)$ is large). In this way, we can choose the data such that (r, s, z) will in fact remain in U up to the time of blow-up. This completes the sketch of the proof.

References:

- [1] L. Boltzmann, Zur Theorie der elastischen Nachwirkung, Ann. Phys. 7 (1876), Ergänzungsband, 624-654.
- [2] B. D. Coleman, M. E. Gurtin and I. R. Herrera, Waves in materials with memory, Arch. Rat. Mech. Anal. 19 (1965), 1-19; B. D. Coleman and M. E. Gurtin, *ibid*, 239-265.
- [3] C. M. Dafermos, Dissipation in materials with memory, in: J. A. Nohel, M. Renardy and A. S. Lodge (eds.), Viscoelasticity and Rheology, Academic Press, to appear.
- [4] C. M. Dafermos, Development of singularities in the motion of materials with fading memory, Arch. Rat. Mech. Anal., to appear.
- [5] C. M. Dafermos and J. A. Nohel, Energy methods for nonlinear, hyperbolic Volterra integrodifferential equations, Comm. PDE 4 (1979), 219-278.
- [6] C. M. Dafermos and J. A. Nohel, A nonlinear hyperbolic Volterra equation in viscoelasticity, Amer. J. Math., Supplement (1981), 87-116.
- [7] H. Hattori, Breakdown of smooth solutions in dissipative nonlinear hyperbolic equations, Q. Appl. Math. 40 (1982/83), 113-127.
- [8] W. J. Hrusa, Global existence and asymptotic stability for a semilinear hyperbolic Volterra equation with large initial data, SIAM J. Math. Anal. 16 (1985), 110-134.
- [9] W. J. Hrusa and J. A. Nohel, Global existence and asymptotics in one-dimensional nonlinear viscoelasticity, in: P. G. Ciarlet and M. Roseau (eds.), Trends and Applications of Pure Mathematics to Mechanics, Springer Lecture Notes in Physics 195 (1984), 165-187.
- [10] W. J. Hrusa and J. A. Nohel, The Cauchy problem in one-dimensional nonlinear viscoelasticity, J. Diff. Eq., to appear.
- [11] W. J. Hrusa and M. Renardy, On a class of quasilinear partial integro-differential equations with singular kernels, J. Diff. Eq., to appear.
- [12] F. John, Formation of singularities in one-dimensional nonlinear wave propagation, Comm. Pure Appl. Math. 27 (1974), 377-405.
- [13] S. Klainerman and A. Majda, Formation of singularities for wave equations including the nonlinear vibrating string, Comm. Pure Appl. Math. 33 (1980), 241-263.
- [14] P. D. Lax, Development of singularities of solutions of nonlinear hyperbolic partial differential equations, J. Math. Phys. 5 (1964), 611-613.
- [15] R. C. MacCamy, A model for one-dimensional nonlinear viscoelasticity, Q. Appl. Math. 35 (1977), 21-33.

- [16] R. Malek-Madani and J. A. Nohel, Formation of singularities for a conservation law with memory, *SIAM J. Math. Anal.* 16 (1985), 530-540.
- [17] P. A. Markowich and M. Renardy, Lax-Wendroff methods for hyperbolic history value problems, *SIAM J. Num. Anal.* 21 (1984), 24-51; Corrigendum, *SIAM J. Num. Anal.* 22 (1985), 204.
- [18] M. Renardy, Recent developments and open problems in the mathematical theory of viscoelasticity, in: J. A. Nohel, M. Renardy and A. S. Lodge (eds.), Viscoelasticity and Rheology, Academic Press, to appear.
- [19] M. Slemrod, Instability of steady shearing flows in a nonlinear viscoelastic fluids, *Arch. Rat. Mech. Anal.* 68 (1978), 211-225.
- [20] M. Slemrod, Appendix: Breakdown of smooth shearing flow in viscoelastic fluids for two constitutive relations: the vortex sheet vs. the vortex shock, in: D. D. Joseph, *Hyperbolic phenomena in the flow of viscoelastic fluids*, to appear in: J. A. Nohel, M. Renardy and A. S. Lodge (eds.), Viscoelasticity and Rheology, Academic Press.

A FAST ALGORITHM FOR NON-NEWTONIAN FLOW

David S. Malkus
Mathematics Research Center
University of Wisconsin - Madison
Madison, WI 53705

ABSTRACT. The goals of the project described here are twofold: First, to turn an existing pilot algorithm for the steady flow of non-Newtonian memory fluids into a robust and efficient algorithm. Second, render enhancements of the method's current capabilities computationally feasible. Such enhancements include fully coupled thermal dependence, material compressibility, and free surface flows. The pilot algorithm is a finite element method whose novelty lies in its computation of the stress field in a nonlinear iteration scheme. The stress at a point is a non-local functional of the current velocity iterate, and the pilot method has demonstrated the feasibility of reliable computation with such constitutive equations. Before the method can take its place as a reliable scientific and engineering tool, intensive effort must be made to reduce the computational cost in the manner described here.

I. VISCOELASTIC FLUIDS. The following equations are solved numerically, using the finite element method [1 - 4]: The equations of steady motion,

$$\nabla \cdot \boldsymbol{\sigma} + \mathbf{f} = \rho(\mathbf{u} \cdot \nabla)\mathbf{u} \quad (1)$$

where \mathbf{u} is the velocity field, $\boldsymbol{\sigma}$ the stress tensor, \mathbf{f} a body force, and ρ the density. The equation of continuity for an incompressible fluid is

$$\nabla \cdot \mathbf{u} = 0 \quad (2)$$

For non-Newtonian fluids, the crucial equation is the constitutive equation,

$$\boldsymbol{\sigma} = -p\mathbf{I} + 2\mu(0)R\dot{\mathbf{e}} + (1 - R)\boldsymbol{\sigma}' \quad (3)$$

where p is an isotropic contribution to the stress, $\mu(0)$ is a zero-shear viscosity, R is a ratio of a retardation time, Λ , to a retardation time, T , and $\boldsymbol{\sigma}'$ is an extra stress tensor. The ratio, R , and its complement determine the proportion of the stress which is Newtonian — and usually is the result of a Newtonian solvent — and the complementary proportion from the extra stress — usually due to long-chain molecules (such as polymers) dissolved in the Newtonian solvent.

There are many proposed forms for the extra-stress tensor: there are two basic categories: the differential and integral models [1]. Here we shall only be concerned with the integral form.

$$\begin{aligned} \boldsymbol{\sigma}' &= \frac{\mu_0}{T} \sum_{l=1}^M \int_{-\infty}^t \mathbf{S}_0^{(l)}(\tau) m_l(\tau) d\tau \\ m_l(\tau) &= T^{-1} \sum_{k=1}^N G_k^{(l)} p_k^{(l)} \exp\left(-\frac{\tau p_k^{(l)}}{T}\right) \end{aligned} \quad (4)$$

where μ_0 is a constant determining $\mu(0)$, and $S_0^{(l)}$ is a strain measure, measuring the deformation which carried the particle from its position at time τ in the past to the stress evaluation point at the present time, 0. The strain measures are the same kind employed in finite elasticity. The memory functions, m_l , are usually sums of exponentials, each with amplitude determined by the modulus, $G_k^{(l)}$ and decay constant, p_k , which determines the fraction of the basic decay rate, T . Thus a computational method must determine the deformation history of every stress evaluation point required to solve the equations of motion in some approximate way, compute the required strain measure — which is almost always highly nonlinear in its dependence on the velocity field — and then approximate the history integral over an infinite interval. This just computes the stress, and then the stress computation must be imbedded in some iterative scheme to produce an approximate solution to the highly nonlinear equations of motion.

II. SOME FLOWS. In this section we give a brief description of some of the flows to which the current method is being applied. The geometry of these flows is quite simple and the results obtained do not illustrate the real power of the finite element method. It is hoped that the reader will appreciate that the method described here is still very much in the development stage, and that the problems so far investigated by the author and other researchers are intended to isolate the complexity inherent in the non-Newtonian nature of the flow from other possible complications. Nevertheless, there seems to be a good deal of physical interest in the problems pictured here, in spite of their geometric simplicity.

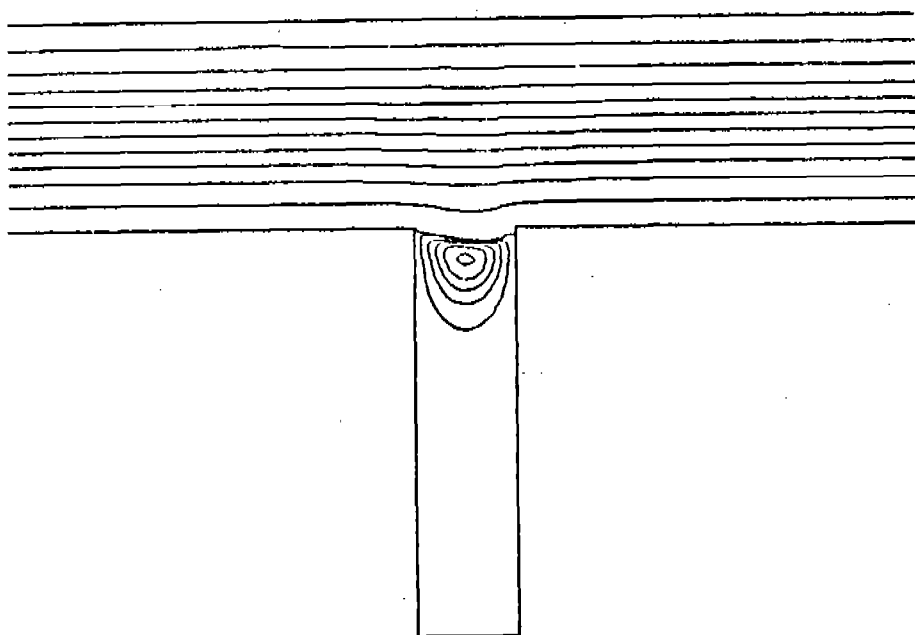


FIGURE 1

Flow over a transverse slot computed by the pilot method on a mesh of 1008 elements.

The first flow is a plane flow over a transverse slot. The streamlines plotted in Figure 1 are taken from a solution computed by the author, using a constitutive equation of his own devising [1] and the mesh of 1008 crossed-triangle macroelements illustrated in ref. 1. The flow is at a "Deborah number" of 4.7 (this can be thought of as a dimensionless shear rate). Flow is from right to left, and undisturbed flow profiles have been imposed at the inflow and outflow. Actually, since there is fluid memory, the inlet condition is that the flow continues forever upstream as undisturbed plane Poiseuille flow. Figure 1 illustrates a characteristic tilt to the vortex in the slot, which is opposite in direction to the tilt observed in Newtonian flows with non-zero Reynolds number [3].

The interest in flows over transverse slots arises from the fact that there seems to be an important relation between the difference between the pressures at top and bottom of the slot and the first normal-stress difference of the fluid in the undisturbed flow [1 - 3]. There seems to be a discrepancy between what the numerical models predict and laboratory experiments measure in such flows, and it is one of the author's highest priorities to resolve this. The results could have important ramifications for devices designed to measure the first normal-stress difference using "hole-pressure" measurements.

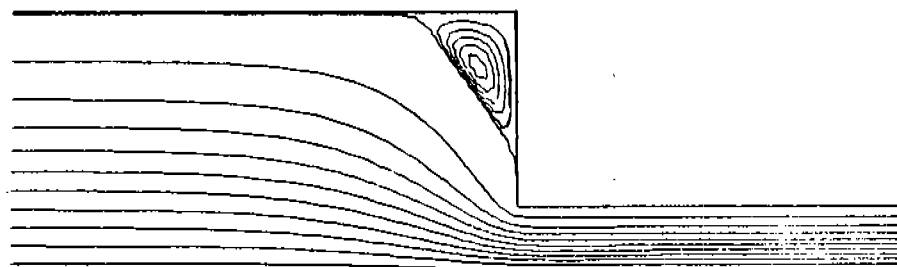


FIGURE 2

Abrupt contraction flow computed by the pilot method with 940 elements.

The next flow is that of flow through an abrupt, planar contraction. Figure 2 pictures such a flow; flow is from left to right so that the fluid is being forced from the larger to the smaller channel. Because symmetry is assumed, the computational domain is only the top half of a channel cross-section. The flow pictured here uses the same fluid model as that of Figure 1, at a slightly lower Deborah number, 3.7. Inflow and outflow conditions

are imposed as before; extreme care is taken to match the flux at inflow and outflow boundaries.

The interest in this flow stems from the fact that some fluids seem to behave quite differently than others in contraction flow. Some fluids, such as polystyrene or high-density polyethylene melts, seem to have relatively smaller "dead-spaces" or recirculation regions at high Deborah number than at low Deborah number, while some branched polymers, such as low-density polyethylene, seem to do quite the opposite, developing recirculation regions emanating from entry which dominate the whole flow-field. The flow pictured in Figure 2 has about the same size recirculation region as a flow of the same fluid at low shear rate. The author is interested in further study of this flow in order to find out what property of the constitutive model is associated with entry vortex behavior. The ability to predict recirculation size is of practical import because fluid trapped inside dead-spaces tends to degrade. It would be useful to be able to determine how much polymer degradation could be expected in a given die as a function of measurable material properties.

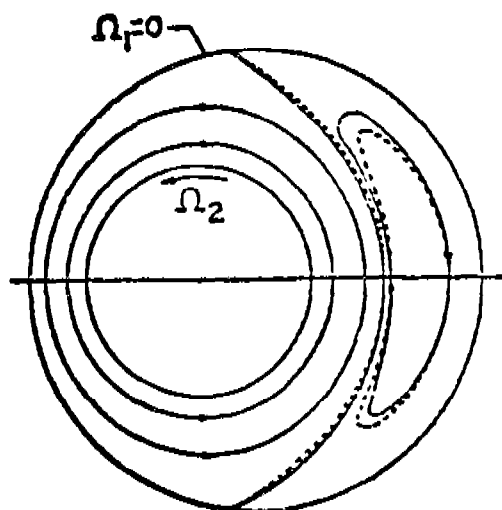


FIGURE 3

Newtonian flow in a plane cross-section of a journal-bearing.

The final flow is one for which the author has as yet no results; that is the flow in a journal-bearing. Figure 3 pictures the cross-section of two eccentrically placed cylinders with fluid between them to lubricate and prevent solid-to-solid contact. There are two important aspects to this flow which differ from the previous two flows: First, there are no inflows and outflows, and second, there are no domain corners to generate singularities in the stress-field.

The author's interest in this problem is, first, just that it is quite different from the other two. It will be interesting to observe the behavior of the numerical method here because it omits two puzzling aspects of memory-fluid problems: history-dependence at inlets, and stress-singularities of unknown character. There is also physical interest in this problem because the effects of fluid elasticity on the load-bearing capacity of the bearing may be beneficial. It would be able to predict load bearing-capacity from measurable material properties, and numerical modelling may help to do so.

III. COMPUTATIONAL METHOD. The strain measure in the integrand of eq. (1) is assumed to be determined by a deformation gradient, $\mathbf{E}_0(\tau)$, just as in nonlinear elasticity. Only in the present case, the deformation gradient is assumed to be computable from a system of linear, non-constant coefficient ordinary differential equations along the path followed by each particle at which the stress is to be evaluated. The usual deformation gradients of large-strain elasticity can also be obtained from special cases of the following evolution equations:

$$\begin{aligned}\dot{\mathbf{x}}(\tau) &= \mathbf{v}[\mathbf{x}(\tau)] \\ \mathbf{x}(0) &= \mathbf{x}_0 \\ \dot{\mathbf{E}}_0(\tau) &= \mathbf{F}(\mathbf{x}(\tau), \nabla \mathbf{v}[\mathbf{x}(\tau)]) \mathbf{E}_0(\tau) \\ \mathbf{E}_0(0) &= \mathbf{I}\end{aligned}\tag{5}$$

The first two sets of equations determine the pathline (streamline) followed by a particle to bring it from its position, \mathbf{x} , at time τ in the past to its present position at the stress evaluation point, \mathbf{x}_0 . To evaluate the integrand of eq. (4), these equations are solved as an initial value problem in reverse time. This determines the non-constant coefficient in the evolution equation for the gradient, which is assumed to be a traceless matrix, \mathbf{F} . The common deformation gradient, $\frac{\partial \mathbf{x}(\tau)}{\partial \mathbf{x}_0}$, is obtained when \mathbf{F} is $\nabla \mathbf{v}$ itself.

The fundamental strategy of the current numerical method is to choose constant strain-rate finite elements: then the evolution equation is a constant-coefficient equation on each element. This strategy is enabled by a basic property of linear ODEs: If we define a deformation gradient, \mathbf{E}_{τ_1} , relative to time τ_1 by

$$\begin{aligned}\dot{\mathbf{E}}_{\tau_1} &= \mathbf{F} \mathbf{E}_{\tau_1} \\ \mathbf{E}_{\tau_1}(\tau_1) &= \mathbf{I}\end{aligned}\tag{6}$$

then the strain relative to the present time, evaluated at any earlier time is give by matrix multiplication:

$$\mathbf{E}_0(\tau) = \mathbf{E}_0(\tau_1) \mathbf{E}_{\tau_1}(\tau)\tag{7}$$

This provides interface conditions between finite elements, so that only constant-coefficient equations need be solved on each element; it turns out that such solutions are known analytically, as is the pathline and transit time along it [1 — 3].

Thus, given an estimate of the solution to the problem in terms of a velocity field, the integrand of eq. (4) can straight-forwardly be computed at each historical time. In the

current method, this is used in conjunction with a specially devised Gaussian quadrature formulas to approximate σ' :

$$\int_{-\infty}^0 S_0(\tau) m(\tau) d\tau \approx \sum_{k=1}^{N_p} \omega_k S_0(\tau_k) \quad (8)$$

With what we have thus far, the stress can be computed in any trial velocity field; to approximate the solution to eqs. (1) and (2), the usual Galerkin procedure can be followed in which the residual of eq. (1) is dotted into a test function, \mathbf{v}^h , drawn from the same space as the trial solutions, and the result is integrated over the problem domain. After integration by parts and replacement of the spatial integral by a numerical integral with points ξ_e and weights θ_e , we get something which looks like

$$\sum_e \theta_e [\sigma' : \nabla \mathbf{v}^h - 2z(\nabla \cdot \mathbf{u}^h)(\nabla \cdot \mathbf{v}^h) + \rho(\mathbf{u} \cdot \nabla) \mathbf{u}^h \cdot \mathbf{v}^h - \mathbf{v}^h \cdot \mathbf{f}](\xi_e) = 0 \quad (9)$$

The pressure term of eq. (3) has been replaced by a penalty term [3] with penalty parameter z ; thus there are no explicit pressure unknowns, and the continuity equation (2) is satisfied to $O(z^{-1})$. Eq. (9) illustrates the $R = 0$ case; for nonzero R , the obvious modification of adding a Newtonian viscous term is made.

The important point to observe about eq. (9) is that to evaluate its residual, it is required to evaluate the stress at the points ξ_e by means already discussed. To complete the method, what is needed is a means of correcting estimates of the discrete solution, based on evaluation of the residual; Newton's method might an example of such a procedure, but, as we shall see, this is not entirely straight-forward. The current algorithm employs the inverse Broyden method [1,2] to solve the discrete nonlinear equations. An important point to be made here is that, regardless of the choice of iterative scheme, the method outlined here is enormously costly in practice, because for a reasonably fine mesh, each evaluation of the stress-field values at the spatial integration points is a potentially formidable computation.

IV. FAST ALGORITHMS. The method outlined in the previous section applies to isothermal, incompressible flows in a fixed spatial domain. These restrictions are not essential; material compressibility and temperature dependence can be handled in very similar fashion if "artificial (historical) time" is introduced, in which either density or temperature are used to change the time variable along the pathlines in such a way that a traceless matrix in the evolution equation is obtained [5]. The transformation to artificial time does not in itself seem to be computationally costly, but these problems involve added levels of complexity to an already complicated solution procedure with additional fields and corresponding equations. The resulting phenomena are likely to be more intricate in detail and more nonlinear in character. A similar observation can be made about free-surface flows; a well-developed methodology exists [6] to solve such problems, which can be directly interfaced with the method outlined here, but this also certain to render the computations more formidable than they now are. With the current algorithm, computations on a mesh which is refined only to the extent which seems to be required to obtain acceptable accuracy, at a shear rate normally occurring in polymer processing, the computation of a steady

solution can take as long as 40 minutes on a Cray 1-A. This must be reduced drastically if the algorithm is to be used routinely in scientific and engineering research, particularly if the more physically realistic enhancements mentioned above are to be added. The remainder of this paper will discuss several approaches to the reduction in computational cost which are currently being implemented or investigated by the author.

Vectorization of Linear Equation Solving. A variety of new computers have the capability of carrying out hardware vector operations; rearranging the computer code in such a way that the compiler can take advantage of this capability can result in substantial savings in computational cost. One part of a typical code where such savings have a good chance of being realized is in the solution of linear equations. Unfortunately, in the current algorithm, it is not expected that this can dramatically reduce the run time. Linear equations are solved in the nonlinear iteration scheme, but this appears to account for a small portion of the computational cost. The major portion of the calculation is carried out at the element level with small arrays or scalar quantities involved in resolving element boundary crossings and accumulating the deformation gradient by small matrix multiplication. Vectorization offers little hope of speeding up these calculations. On the other hand, it is expected that linear equation solving will begin to play a more and more important role with the planned enhancements to the code discussed earlier. The Jacobian terms corresponding to the thermal energy equation are easy to form; likewise the part of the Jacobian associated with inertial terms and the unknown free-surface transformation are easy to deduce. Also, active research is under way aimed at producing the Jacobian terms associated with the non-Newtonian viscous terms (see below).

In short, the future development of the code seems to point in the direction of an algorithm which has a large, unsymmetric, and possibly not banded matrix to factor at each one of dozens of possible iterates. The current iteration method has only one, banded, symmetric, positive-definite matrix to factor at the outset, and a back-substitution at each iteration (the unsymmetric Jacobian contribution of the inertial terms is left to the inverse updating scheme). It therefore seems appropriate to modify the code at the present time to take full advantage of vectorization, in order to make sure the linear equation solving phase remains in the background, as it should.

Adaptive Memory Quadrature. The area which seems to show most promise in reduction of the computational cost is that of the stress calculation at an individual stress evaluation point. There seem to be several possible approaches, the underlying strategy of all of them is to take advantage of the fact that the stresses are being evaluated in what is hoped will be a convergent sequence of velocity iterates. Particularly further along in the sequence, the previous iterate should be able to provide a guide to estimate how much computation is absolutely necessary at the next iteration.

Perhaps the most obvious way to do this is to use the previous iterate to determine what N_p of eq. (8) should be in the next iterate. It is observed that in some flows, very many fewer quadrature points are needed to accurately compute σ' than in other flows. The strategy will be to begin the iterations with a nominal number of points for each stress evaluation point and increase or decrease that number in succeeding iterations, based on

adaptive criteria determined from previous iterations.

Jacobian Approximation. The reason that the present algorithm employs an updating scheme rather than a direct calculation of the Jacobian is that it is not a trivial matter to construct the Jacobian, or even write down a closed form expression for it. It is clear that the stress at a point can depend on velocities far from that point, and therefore the Jacobian of the residual of eq. (9) cannot have the usual finite element band and/or sparsity structure. In ref. 7, an approximation scheme for the Jacobian is proposed. It is not clear at present whether this approximation, some other, or even an exact computation of the Jacobian is best (the latter may be possible to undertake — it is not clear at this time). But the work of ref. 7 shows clearly the complexity involved. The terms of the Jacobian contribution from the extra stress are computed by tracking along streamlines. The resulting Jacobian element matrices are not square: Their column dimension depends on the number of different elements the particle path passes through before the final integration point of eq. (8) is located. It appears that a frontal solution technique is called for in order to handle the resulting global matrix [7].

Pseudo-Dynamic Relaxation. The hope in computing the Jacobian is that the computational cost will be more than returned in improved convergence rate over the inverse Broyden algorithm afforded by Newton's method or modified Newton's method with Broyden updates. But the complexity of the Jacobian calculation is such that this may never be realized, and it is well worth the investigation of other improvements to the iterative solution of the nonlinear equations. One avenue currently being explored is that of "pseudo-dynamic relaxation." The problem is cast as a time-dependent problem and steady solutions are obtained by letting the transient phenomena die out. In the algorithm presented here, the transient behavior does not represent the true dynamics of the non-Newtonian fluid; the stress is computed in the current velocity field as if it had been a steady field for all time, hence the name "pseudo-dynamic." To do otherwise would involve complexities beyond what seems manageable at present, though implementation of the pseudo-dynamic algorithm does open the door for future exploration of true dynamic behavior.

One may easily verify that the steady-states of the pseudo-dynamic algorithm are the same as the steady states of the true dynamic algorithm. The reason for taking the pseudo-dynamic approach is to produce a different kind of steady-state iteration scheme, in which the damping of the high frequency modes in the pseudo-dynamic response can be controlled by choice of time-stepping method and pseudo-time step. The reason that this seems to be a worthwhile avenue to explore is suggested by recent work of Y. Renardy and M. Renardy [8]. They found that with a certain spatial discretization of the linearized operator associated with the equations of motion of a Maxwell fluid in a shearing flow, there were apparently spurious eigenvalues extremely close to the right half-plane, evidently introduced by the discretization. If this were also a consequence of finite element discretization, there could be severe consequences for iterative methods which behave like temporal iteration schemes. It is hoped that by controlling the time step and parameters of the pseudo-dynamic time-stepping method, the damping of the high frequency modes associated with any spurious eigenvalues can be damped to produce nearer monotonic,

more rapid convergence of the resulting iterative method. There is no worry here of damping out interesting transient behavior — the transient behavior is not correct, and all that it is required is that it be damped out as rapidly as possible.

The following algorithm is based on Hughes, Liu, and Brooks predictor-corrector algorithm for the Navier-Stokes equations [9], but with the possibility of more fully implicit inner iterations at each time step:

$$\begin{aligned}
 \mathbf{M}\mathbf{a}_{n+1} + \mathbf{C}\mathbf{v}_{n+1} - \mathbf{N}(\mathbf{v}_{n+1}) + \mathbf{Q}(\mathbf{v}_{n-1}) &= \mathbf{F}_{n+1} \\
 \mathbf{Q} &= \int_{\Omega} \mathbf{B}^T \boldsymbol{\sigma} dV - \mathbf{C}\mathbf{v} \\
 \mathbf{N} &= \text{nonlinear inertial term, excl. time deriv.} \\
 \mathbf{v}_{n+1}^{(0)} &= \mathbf{v}_n + (1 - \gamma)\Delta t \mathbf{a}_n \\
 \mathbf{v}_{n+1}^{(i+1)} &= \mathbf{v}_{n+1}^{(i)} - \mathbf{J}^{-1} \{ [\mathbf{M} + \gamma\Delta t \mathbf{C}] \mathbf{v}_{n+1}^{(i)} + \gamma\Delta t \mathbf{Q}(\mathbf{v}_{n+1}^{(i)}) \\
 &\quad + \gamma\Delta t \mathbf{N}(\mathbf{v}_{n-1}^{(i)}) - \mathbf{M}\mathbf{v}_{n+1}^{(0)} + \gamma\Delta t \mathbf{F}_{n+1} \} \\
 \mathbf{a}_{n+1} &= (\mathbf{v}_{n+1} - \mathbf{v}_{n+1}^{(0)}) / \gamma\Delta t \\
 \mathbf{J} &= \mathbf{M} - \left(\gamma\Delta t \frac{\partial \mathbf{Q}}{\partial \mathbf{v}} \right)_{opt} + \left(\gamma\Delta t \frac{\partial \mathbf{N}}{\partial \mathbf{v}} \right)_{opt}
 \end{aligned} \tag{10}$$

\mathbf{M} is the finite element “mass matrix,” \mathbf{C} the Newtonian viscous and penalty-pressure matrix, and \mathbf{F}_{n+1} the applied force vector at time step $n + 1$. \mathbf{B} is the usual finite element matrix of shape function derivatives and $\boldsymbol{\sigma}$ is the stress, computed in $\mathbf{v}_{n+1}^{(i-1)}$ as described in previous sections: \mathbf{v}_{n+1} without the superscript of inner iteration is the “fully converged” result of inner iteration at time-level $n + 1$. Choice of the number of inner iterations is open, so that \mathbf{v}_{n+1} could result from just one correction cycle, or many. An important aspect of eq. (10) is found in those terms labelled by $(\cdot)_{opt}$; a fully implicit treatment would employ exact Jacobian terms here. At the other extreme is Hughes, Liu, and Brooks method: they use \mathbf{C} to approximate both of these terms and do only one inner iteration. The present non-Newtonian implementation uses \mathbf{C} initially, updated by the inverse Broyden method during a number of inner iterations. If it proves to be effective, Newton or modified Newton/Broyden iterations could be employed in the inner iterations. It is instructive to note that the direct steady-state Broyden algorithm mentioned earlier is obtained as a special case of eq. (10) with $\gamma = 1$ and an infinite time step.

V. SUMMARY. A pilot numerical method for the computation of solution to memory fluid flow problems has been described. This method has shown that such computations are feasible but extremely costly. More reasonable physical assumptions than those of isothermal, incompressible flow in a fixed domain are on the near horizon but are bound to increase the computational cost. A number of ways of improving the computational performance of the algorithm have been proposed here and are in the implementation stage. These improvements will go together to make what the author refers to as “a fast algorithm for non-Newtonian flow.” It is hoped that this fast algorithm can transform the method described here from pilot code to useful computational tool for the investigation of problems in viscoelasticity and rheology.

Acknowledgements: The numerical techniques for integral constitutive equations described here were developed jointly by the author and B. Bernstein (Dept. Mathematics, I. I. T.). Development of the fast algorithm for non-Newtonian flow is sponsored by the United States Air Force, under Grant No. 84-NM-399. The author also acknowledges the support of the National Science Foundation (Grants MCS 79-03542, 81-02089, and 83-01433, and the Supercomputer Initiative, which provided access to the University of Minnesota Cray - 1A). At the Mathematics Research Center, University of Wisconsin - Madison, the author's research is sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

REFERENCES.

1. D. S. Malkus, finite element methods for viscoelastic flow. MRC Technical Summary Report No. 2812, Mathematics Research Center, University of Wisconsin, Madison, WI (1985).
2. D. S. Malkus and B. Bernstein, Flow of a Curtiss-Bird fluid over a transverse slot using the finite element drift-function method, *J. Non-Newtonian Fluid Mechs.* 16, 77(1984).
3. B. Bernstein, D. S. Malkus, and E. T. Olsen, A finite element for incompressible plane flows of fluids with memory, *Int. J. Numer. Meths. Fluids* 5, 43(1985).
4. B. Bernstein, M. K. Kadivar, and D. S. Malkus, Steady flows of memory fluids with finite elements: Two test problems, *Comp. Meths. Appl. Mechs. Eng.* 27, 279 (1981).
5. B. Bernstein, private communication.
6. H. S. Kheshgi and L. E. Scriven, Penalty finite element analysis of time-dependent two-dimensional free surface film flows, in *Finite Element Flow Analysis*, Ed: T. Kawai, University of Tokyo Press, Tokyo (1982).
7. G. Zazi, Ph. D. Thesis, Illinois Institute of Technology, Chicago, IL (1985).
8. Y. Renardy and M. Renardy, Trans. 3rd Army Conf. on Appl. Math. and Computing, this volume.
9. T. J. R. Hughes, W. K. Liu, and A. Brooks, Finite element analysis of incompressible viscous flows by the penalty function formulation, *J. Comp. Phys.* 30, 1 (1979).

WAVE CURVES FOR THE RIEMANN PROBLEM OF PLANE WAVES
IN SIMPLE ISOTROPIC ELASTIC SOLIDS *

Zhijing Tang and T. C. T. Ting
Department of Civil Engineering, Mechanics and Metallurgy
University of Illinois at Chicago
Chicago, Illinois 60680

ABSTRACT A modified Riemann problem in which the initial and boundary conditions are constants are considered for plane waves in a half space occupied by a simple elastic solid. The governing quasilinear differential equations are a system of hyperbolic conservation laws which possesses three wave speeds $c_1 > c_2 > c_3$. The system is genuinely nonlinear with respect to c_1 and c_3 and linearly degenerate with respect to c_2 . Thus it is sufficient to study a two-wave speed system with c_1 and c_3 . Wave curves for simple waves and shock waves are used to construct the solution. Second order hyperelastic materials which contain four material constants are considered and the solution in the form of wave curves are obtained for all possible combinations of initial and boundary conditions. With a proper nondimensionalization, the wave curves depend only on one material parameter k . The solutions are thermodynamically correct because the entropy effects does not come into picture until the third order in stresses are included in the constitutive laws. The two-wave speed system have one umbilic point at which $c_1 = c_3$ and hence the system is not totally hyperbolic (or not strictly hyperbolic). Several interesting and unexpected results are obtained due to the existence of the umbilic point. In one example, we find that a shock wave satisfies Lax stability condition for a V_1 shock as well as a V_3 shock. In another, a shock wave which involves only one stress component does not satisfy Lax stability condition for either V_1 shock or V_3 shock. However, it satisfies Lax stability condition if we consider it under the context of one-wave speed system. Finally we consider the effects on the solution when the third order terms are included. We show that although the entropy affects the shock wave solution, it does not appear in the simple wave solution until the fourth order terms are included. With the third order terms, there are in general two umbilic points, one of which may be an umbilic line.

EXTENDED SUMMARY The Riemann problem is a special case of the Cauchy problem in which the initial value is a constant for $X > 0$ and a different constant for $X < 0$. Physically, this may represent the problem of a fluid in a tube with different initial pressures on both sides of a diaphragm located at $X = 0$. The solution to the Riemann problem yields the fluid motion after the diaphragm is ruptured. The problem we will study here is the plane waves in an elastic half space $X = X_1 \geq 0$ which is prestressed at time $t = 0$. For $t > 0$, a constant traction is applied at the boundary $X = 0$. This is a "modified" Riemann

* This work is supported by the U. S. Army Research Office

problem.

The governing quasilinear partial differential equations for the problem are a system of hyperbolic conservation laws. Extensive coverages of the subject can be found in Courant and Friedrichs (1948), Jeffrey (1976), Lax (1957, 1973) and Dafermos (1983), for example. Inherited with the solution to the quasi-linear system is the weak solution or the shock wave in which the solution is discontinuous. The appearance of shock waves in the solution is a general rule rather than an exception.

Plane finite amplitude waves in simple elastic solids have been studied by many investigators. Chu (1964) and Collins (1966) considered incompressible materials, Bland (1964a,b, 1965) and Davison (1966) studied compressible elastic solids while Howard (1966) investigated waves in transversely isotropic materials. For the Riemann problem, the solution in general consists of simple waves and/or shock waves (Lax 1957, Liu 1975, Smoller 1969b). The crux of the problem is in the determination of the correct sequence of simple waves and shock waves to satisfy the initial and boundary conditions. Most investigators used a semi-inverse approach, i.e. one assumed a combination of simple waves and shock waves to see what initial and boundary conditions were satisfied. A direct approach using the wave curves for simple waves and shock waves was employed by Li and Ting (1983). The idea has been used earlier by Clifton (1966) and Ting and Nan (1969) where the wave curves are called "stress paths". However, the problems studied by them were for elastic-plastic materials in which the elastic part was linear and hence the wave curves for shock waves were straight lines. In the paper by Li and Ting (1983), the second order hyperelastic material was studied for a special case in which the material was initially stress free. In the present paper we extend their problem to include arbitrary initial condition. We also consider the effects on the solution of including the third order terms in the constitutive equations.

In Chapter II, the basic equations for the problem are developed. The material is assumed to be isotropic simple elastic solids. Following Li and Ting (1983), we use the stresses instead of the deformation gradients as the dependent variables. This is due to the fact that it is more natural to prescribe stress rather than deformation gradient (or strain) as the initial and boundary conditions. With the assumption of plane waves in isotropic simple elastic solids, the deformation gradients are functions of σ and τ^2 only, where σ and τ are, respectively, the normal and shear stress on a $X = X_1 = \text{constant}$ plane. The angle the shear stress makes with respect to X_2 axis is denoted by θ . Since the basic solutions to the Riemann problem involve simple waves and shock waves, we discuss simple wave solutions in Chapter III and shock wave solutions in Chapter IV.

A simple wave solution represents a wave fan on the (X,t) plane in which the stresses are constants along the straight lines passing through the origin. As the slope of the straight lines varies, the stresses trace a curve in the stress space which is called the simple wave curve. There are three simple wave curves associated with the three wave speeds $c_1 \geq c_2 \geq c_3$. Using the cylindrical coordinate

system (σ, τ, θ) for the stress space, it is shown that the simple wave curves associated with c_1 and c_3 lie on a radial plane ($\theta = \text{constant}$) and hence are plane polarized. The simple wave curve associated with c_2 is a circle ($\sigma = \text{constant}$ and $\tau = \text{constant}$) and hence is circularly polarized. These results agree with that of Bland (1965) except that Bland used deformation gradients as the dependent variables. We then use the second order hyperelastic materials to study the geometry of simple wave curves. There are four material constants but the simple wave curves depend only on one non-dimensional parameter k . Depending on the value of k , there are four different simple wave curves.

In Chapter IV we study the shock wave solutions. For a given stress state in front of the shock, the admissible stress state behind the shock depends on the shock wave speed V . As V varies, the stress state behind the shock traces a curve in stress space which is called the shock wave curve. There are three possible shock wave curves associated with the three shock wave speeds V_1 , V_2 and V_3 . Like simple wave curves, shock wave curves associated with V_1 and V_3 are plane polarized and the one associated with V_2 is circularly polarized. The latter is identical to the simple wave curve associated with c_2 since the governing differential equations are linearly degenerate with respect to c_2 (Lax 1957).

A shock wave solution satisfying the Rankine-Hugoniot jump conditions is not necessarily admissible unless it also satisfies the Lax stability condition (Lax 1957)

$$c_i(B) \leq V_i(B, A) \leq c_i(A) , \quad (1.1)$$

where B and A denote, respectively, the stress state in front of (or Before) and behind (or After) the shock wave. Without Lax stability condition the solution may not be unique. A simple example of non-uniqueness is provided by Hopf's equation (Hopf 1950, Witham 1974). For a large amplitude shock, Lax stability condition is necessary but not sufficient. A more discriminating condition was proposed by Liu (1974). Let P be any point between A and B on the shock curve. The Liu admissibility condition reads

$$V_i(B, P) \leq V_i(B, A) . \quad (1.2a)$$

Also, if we define the "reversed" shock wave curve the locus of the stress state B in front of the shock for a fixed stress state A behind the shock, and if Q is any point on the reversed shock wave curve between B and A , then the Liu admissibility also stipulates that

$$V_i(B, A) \leq V_i(Q, A) . \quad (1.2b)$$

In general the shock wave curve and the reversed shock wave curve are different. In any case we have

$$V_i(B,B) = c_i(B) \quad \text{and} \quad V_i(A,A) = c_i(A) , \quad (1.3)$$

and (1.1) is a special case of (1.2a,b). As in Chapter III, we use the second order hyperelastic material to study the geometry of shock wave curves. Again, the shock wave curves are found to depend only on one parameter k .

By neglecting the entropy in the second order materials the analysis is thermodynamically correct because the effect of entropy on the shock discontinuity is in the third order of strain (or stress) and, with the adiabatic approximation, the entropy is a constant through a simple wave. We therefore study the solutions to the Riemann problem for second order hyperelastic materials in Chapter V. This chapter represents the main work of this paper. We consider all possible combinations of initial value and boundary value, both of which are constants. In view of the fact that the c_2 simple wave (which is also the V_2 shock wave) is linearly degenerate, there is no loss of generality in ignoring the c_2 simple wave. Therefore all we have to consider are simple wave curves associated with c_1 and c_3 and shock wave curves associated with V_1 and V_3 . These curves lie on the (σ, τ) plane. Consequently, the problem is reduced to a two-wave speed system instead of a three-wave speed system. If $c_1 \neq c_3$, the system is said to be "totally hyperbolic" or "strictly hyperbolic" (Courant and Hilbert 1962, Lax 1957). In our case the system is totally hyperbolic everywhere except at the point $(\sigma, \tau) = (\sigma^*, 0)$ at which $c_1 = c_3$. This point is called the umbilic point (Shearer et al, 1985).

The existence of the umbilic point leads to the following interesting and unexpected results, some of which have also been found by Schaefer and Shearer (1985) in the problem of oil recovery.

i) For the hyperbolic materials, the simple wave curves associated with different wave speeds are orthogonal to each other. Hence there are two wave curves through each point on the (σ, τ) plane. At the umbilic point, however, the simple wave curves may not be orthogonal to each other (see Figs.1-6). Moreover, there may be infinitely many wave curves passing through the umbilic point (Fig.2. See also Ting 1973)).

ii) For the reduced two-wave speed system, there are in general two wave fans in the solution. The wave fan may be a simple wave, a shock wave or a composite wave in which a shock wave is in contact with a simple wave of the same family (Fig.8). However, when the wave curve passes through the umbilic point, one may have three or even four wave fans. (See wave pattern 10 and 11 of Fig.8.)

iii) From a given point other than the umbilic point on the (σ, τ) plane, one can draw two simple wave curves and two shock wave curves. They are orthogonal at the starting point. As one follows one of the wave curves, the curve may intersect with the other wave curve starting from the same point (Fig.9b). This causes the solution to depend discontinuously on the boundary condition. (See also Li and Ting 1983.)

iv) For a large amplitude shock, we have an example in which the

shock wave speed satisfies Lax stability condition (1.1) for both $i = 1$ and $i = 3$ (see (5.6c)). Thus the shock has the double role of being a V_1 shock as well as a V_3 shock.

v) In another example we have a situation in which Lax stability condition (1.1) is not satisfied for either $i = 1$ or $i = 3$ (see (5.9c)). We have therefore no solutions which satisfy Lax stability condition. However, for the example concerned the shock wave curve is along the σ -axis and hence involves only one stress component. Considering the wave motion with one stress component one would obtain a one-wave speed system. Under the one-wave speed system the shock concerned satisfies the Lax stability condition. We have therefore a paradox in which a shock is stable under the one-wave speed system but unstable under a two-wave speed system.

In the last chapter we consider the effects of including the third order terms in the constitutive equations. As pointed out by Bland (1969), we can no longer ignore the entropy since the effect of entropy in a shock wave is of third order in strain and hence in stress. For a special third order hyperelastic material, we show that the entropy jump across the shock is positive for the solutions obtained in Chapter V. A slightly more general third order material is then used to study the effects of third order terms on the geometry of shock wave curves. Finally, we consider the general third order hyperelastic materials. We show that there may be as many as three umbilic points on the (σ, τ) plane, one of which may be an umbilic line. This is interesting since it is not common to have an umbilic line.

REFERENCES

- Bland, D. R. 1964a On shock waves in hyperelastic media. IUTAM Symp. on Second-order Effects in Elasticity, Plasticity and Fluid Dynamics, 93-108
- Bland, D. R. 1964b Dilatational waves and shocks in large displacement isotropic dynamic elasticity. J. Mech. Physics Solids 12, 245-267
- Bland, D. R. 1965 Plane isentropic large displacement simple waves in a compressible elastic solid. Z. Angew. Math. Phys. 16, 752-769
- Bland, D. R. 1969 Nonlinear Dynamic Elasticity. Blaisdell Pub. Co.
- Chu, Bao-Teh 1964 Finite amplitude waves in incompressible perfect elastic materials. J. Mech. Phys. Solids 12, 45-57
- Clifton, R. J. 1966 An analysis of combined longitudinal and torsional plastic waves in a thin walled tube. Proc. 5th U. S. Nat. Congress Appl. Mech. ASME, New York, 465-480
- Collins, W. D. 1966 One dimensional nonlinear wave propagation in incompressible elastic materials. Q. J. Mech. Appl. Math. 19, 259-328
- Courant, R. and K. Friedrichs 1948 Supersonic Flow and Shock Waves. New York, Wiley-Interscience.

- Courant, R. and Hilbert, D. 1962 Methods of Mathematical Physics, Vol. II. Interscience, New York.
- Dafermos, C. M. 1973 Solution of the Riemann problem for a class of hyperbolic systems of conservation laws by the viscosity method. Arch. Rat. Mech. Anal. 52 , 1-9
- Dafermos, C. M. 1974 Quasilinear hyperbolic systems that result from conservation laws. in Nonlinear Waves, Ed. by S. Leibovich and A. R. SeeBass, Cornell Univ. press, 82-102
- Dafermos, C. M. 1983 Hyperbolic systems of conservation laws. Brown Univ. Rept. LCDS #83-5
- Davison, L. 1966 Propagation of plane waves of finite amplitude in elastic solids. Journal of Mechanics and Physics Solids, 14 , 249-270
- Hopf, E. 1950 The partial differential equation $u_t + uu_x = \mu u_{xx}$. Comm. Pure Appl. Math. 3 , 201-230
- Howard, I. C. 1966 Finite simple waves in a compressible transversely isotropic elastic solid. Q. J. Mech. Appl. Math. 19 , 329-341
- Jeffrey, A. 1976 Quasilinear Hyperbolic systems and Waves. Pitman Pub.
- Jeffrey, A. and T. Taniuti 1964 Nonlinear wave propagation with applications to physics and magneto-dynamics. New York, 1964, Academic press.
- Lax, P. D. 1957 Hyperbolic systems of conservation laws II. Comm. Pure Appl. Math. 10 , 537-566
- Lax, P. D. 1973 Hyperbolic systems of conservation laws and the mathematical theory of shock waves. SIAM Publication.
- Li, Y. and T.C.T. Ting 1983 Plane waves in simple elastic solids and discontinuous dependence of solution on boundary conditions. Int. J. Solids Structures, 19 , 989-1008
- Liu, T.-P. 1974 The Riemann problem for general 2X2 conservation laws. Trans. Amer. Math. Soc. 199 , 89-112
- Liu, T.-P. 1975 The Riemann problem for general systems of conservation laws. J. Differential Equations 18 , 218-234
- Schaeffer, D. G. and M. Shearer 1985 The classification of 2X2 systems of non-strictly hyperbolic conservation laws, with application to oil recovery. Comm. Pure Appl. Math. To appear.
- Shearer, M., D. G. Schaeffer, D. Marchesin and P. J. Pese-Leme, 1985 Solution of the Riemann problem for a prototype 2X2 system of non-strictly hyperbolic conservation laws. To appear.
- Smoller, J. A. 1969 On the solution of the Riemann problem with general

step data for an extended class of hyperbolic systems. Michigan Math. J. 16 , 201-210

Ting, T. C. T. 1973 On wave propagation problems in which $c_f = c_s = c_2$ occurs. Q. Appl. Math. 31 , 275-286

Ting, T. C. T. and N. Nan 1969 Plane waves due to combined compressive and shear stresses in a half-space. J. Appl. Mech. 36 , 189-197

Truesdell, C. and W. Noll 1965 The nonlinear field theories of mechanics. Handbuch der physik, Vol. III/3. Berlin, 1965, Springer-Verlag.

Witham, G. B. 1974 Linear and Nonlinear Waves. Wiley-Interscience.

Wright, T. W. 1983 Private communications.

A COMPARISON BETWEEN VECTOR AND TENSOR TRANSFORMATIONS,

AN APPLICATION IN CONTINUUM MECHANICS

M. N. L. NARASIMHAN* and EDWARD A. SAIBEL
Engineering Sciences Division, U.S. Army Research Office,
P. O. Box 12211, Research Triangle Park, N.C. 27709-2211

ABSTRACT. Boundary value problems involving a conformal transformation of simply connected regions when formulated by two commonly used methods, one using a vector approach and the other, a more general tensor approach, reveal an apparent disparity, the resolution of which leads to a condition which must be satisfied by the Jacobian.

As illustrations, (1) the flow of a viscous incompressible fluid in an eccentric annulus and, (2) the torsion of a hollow shaft with a similar geometry, are considered and the conditions to be satisfied by the Jacobians are obtained.

I. INTRODUCTION. Invariant formulations of physical laws and their mathematical modeling require the use of vector and/or tensor approaches. By a vector approach is meant, in the present context, the use of traditional vector calculus operations, typical under a Euclidean-space setting, without the explicit use of covariant differentiation and the Christoffel symbols, which characteristically belong to the domain of tensor calculus. By a tensor approach is meant the use of operations of covariant and intrinsic differentiations which explicitly involve the Christoffel symbols defined in a Riemannian space. These Christoffel symbols are geometry-dependent quantities which, in general, are not tensors themselves except under the affine group of transformations. Also, these symbols identically vanish in the case of Cartesian

*Permanent Address
Department of Mathematics
Oregon State University
Corvallis, Oregon 97331

tensors since the latter are always defined relative to a Cartesian frame of reference. As a consequence of the use of the Riemannian space in the tensor approach, as opposed to that of the Euclidean space, the tensor derivatives which occur in the gradient, divergence and curl operations are involved with a more elaborate geometry acquired through the Christoffel symbols than the geometry involved in the various vector operations. It is the purpose here to examine more closely the consequences of this geometric structure of the tensor approach, which produces extra terms. However, these extra terms must vanish since the vector and tensor approaches must both produce the same outcome. It is shown in this paper that when the above techniques are applied to a physical problem involving a conformal mapping of a simply connected region occupied by a material, there would result a useful geometric condition on the Jacobian of the conformal transformation. In physical problems involving materials with their complicated geometry requiring the use of curvilinear coordinate systems, the tensor approach results in numerous terms whose occurrence poses considerable complexity. The above-mentioned constraining condition on the Jacobian helps readily identify those terms which must vanish, thus resulting in a substantial reduction in the complexity of the problem.

We present some applications of the above concept to both fluid and solid mechanics.

II. FORMULATION

Consider a one-to-one transformation defined by

$$\begin{aligned} z &= F(\zeta), & x^3 &= \lambda^3, \\ z^1 &= x^1 + ix^2, & \zeta &= \lambda^1 + i\lambda^2, & i &= \sqrt{-1}, \end{aligned} \quad (2.1)$$

where

$$x^k, \lambda^k \in \mathbb{R}^3, \quad k = 1, 2, 3,$$

are triplets of real numbers representing two rectangular Cartesian systems.

Let the first of the equations in (2.1) represent a conformal transformation

from the x^1x^2 plane to the $\lambda^1\lambda^2$ plane through the analytic

function F defined over a simply connected open set D such that its derivative

$F'(\zeta) \neq 0$ in D . The Jacobian of the above conformal transformation and

its inverse are, respectively,

$$\begin{aligned} J &= \partial(\lambda^1, \lambda^2) / \partial(x^1, x^2) \\ &= (\partial\lambda^1 / \partial x^1)(\partial\lambda^2 / \partial x^2) - (\partial\lambda^1 / \partial x^2)(\partial\lambda^2 / \partial x^1) > 0, \end{aligned} \quad (2.2)$$

$$\begin{aligned} J^{-1} &= \partial(x^1, x^2) / \partial(\lambda^1, \lambda^2) = (\partial x^1 / \partial \lambda^1)(\partial x^2 / \partial \lambda^2) - (\partial x^1 / \partial \lambda^2)(\partial x^2 / \partial \lambda^1) \\ &= (\partial x^1 / \partial \lambda^1)^2 + (\partial x^2 / \partial \lambda^1)^2 > 0, \end{aligned} \quad (2.3)$$

where the last step in (2.3) follows from the previous one by the use of the Cauchy-Riemann equations

$$\partial x^1 / \partial \lambda^1 = \partial x^2 / \partial \lambda^2, \quad \partial x^1 / \partial \lambda^2 = -\partial x^2 / \partial \lambda^1. \quad (2.4)$$

The line element ds in the $x^1 x^2$ space is related to that in the $\lambda^1 \lambda^2$ space by the following.

$$\begin{aligned} ds^2 &= (dx^1)^2 + (dx^2)^2 \\ &= J^{-1} [(d\lambda^1)^2 + (d\lambda^2)^2] = g_{k\ell} d\lambda^k d\lambda^\ell, \end{aligned} \quad (2.5)$$

where in the last step in (2.5), summation over repeated indices is implied and we shall use this summation convention in the sequel. Thus from (2.5), we find that the metric tensor $g_{k\ell}$ and its reciprocal $g^{k\ell}$ are given by

$$g_{11} = \frac{1}{g_{11}} = J^{-1} = g_{22} = \frac{1}{g_{22}}, \quad g^{k\ell} = 0, \quad g_{k\ell} = 0, \quad \text{for } k \neq \ell. \quad (2.6)$$

Next introduce over the region D , a continuously differentiable, symmetric, second rank, linear tensor-valued function \underline{t} of another tensor-valued function \underline{d} with similar properties, which in turn is a linear function of the gradient of a continuously differentiable vector field \underline{v} as follows.

$$\underline{t}(\lambda^1, \lambda^2) = (\alpha_0 \text{tr} \underline{d}) \underline{I} + 2\alpha_1 \underline{d} \quad (2.7)$$

$$\underline{d}(\lambda^1, \lambda^2) = (1/2)[(\underline{v}\underline{v}) + (\underline{v}\underline{v})^T] \quad (2.8)$$

where \underline{I} = identity tensor, α_0 and α_1 , are scalars independent of the space variables λ^1 and λ^2 , and the superscript T over a tensor denotes its transpose.

Eqs. (2.7) and (2.8) can alternately be expressed in the component notation for the given orthogonal coordinate system as

$$\begin{aligned} t^{k\ell} &= (\alpha_0 \text{tr} \underline{d}) g^{k\ell} + 2\alpha_1 d^{k\ell} \\ &= \alpha_0 v^m{}_{;m} g^{k\ell} + 2\alpha_1 d^{k\ell}, \end{aligned} \quad (2.9)$$

$$\begin{aligned} d^{k\ell} &= d_{mn} g^{mk} g^{n\ell} \\ &= (1/2)(v^k{}_{;\ell} g^{\ell\ell} + v^\ell{}_{;k} g^{kk}) \end{aligned} \quad (2.10)$$

where the semi-colon followed by an index represents the covariant differentiation with respect to space variables and an underscore below an index indicates the suspension of summation over that index. For example, in the $\lambda^1 \lambda^2$ space,

$$v^k_{;l} = \frac{\partial v^k}{\partial \lambda^l} + v^m \{^k_{ml}\}, \quad k, l, m = 1, 2, \dots \quad (2.11)$$

$$d^{12} = (1/2)(v^1_{;2} g^{22} + v^2_{;1} g^{11}) . \quad (2.12)$$

The expression $\{^k_{ml}\}$ in (2.11) denotes the Christoffel symbol of the second kind [SYNGE and SCHILD (1949)] which is related to that of the first kind, $[lm, n]$, and is defined below.

$$\begin{aligned} \{^k_{lm}\} &= \{^k_{ml}\} = g^{kn} [lm, n] \\ &= (1/2) g^{kn} \left(\frac{\partial g_{ln}}{\partial \lambda^m} + \frac{\partial g_{mn}}{\partial \lambda^l} - \frac{\partial g_{lm}}{\partial \lambda^n} \right) . \end{aligned} \quad (2.13)$$

The last step on the right-hand side of (2.9)) follows immediately by taking the trace of (2.8) and noting that

$$\text{tr} \underline{d} = \underline{v} \cdot \underline{v} = v^m_{;m} . \quad (2.14)$$

One of the principal objectives of the present investigation is to compute the divergence of \underline{t} , by two independent methods, vectorial and tensorial, in order to derive the condition to be satisfied by the Jacobian of the conformal mapping (2.1).

The divergence of (2.7) is given by

$$\nabla \cdot \underline{t} = \alpha_0 \nabla(\nabla \cdot \underline{v}) \underline{I} + 2\alpha_1 \nabla \cdot \underline{d} , \quad (2.15)$$

where $\nabla \cdot \underline{d}$ is obtained from (2.8):

$$\begin{aligned} \nabla \cdot \underline{d} &= (1/2) [\nabla(\nabla \cdot \underline{v}) + \nabla^2 \underline{v}] \\ &= (1/2) [2\nabla(\nabla \cdot \underline{v}) - \nabla \times \nabla \times \underline{v}] , \end{aligned} \quad (2.16)$$

so that (2.15) becomes

$$\nabla \cdot \underline{t} = \alpha_0 \nabla(\nabla \cdot \underline{v}) \underline{I} + \alpha_1 [2\nabla(\nabla \cdot \underline{v}) - \nabla \times \nabla \times \underline{v}] . \quad (2.17)$$

The physical components of $\nabla \cdot \underline{t}$ in the given orthogonal coordinate system $\lambda^1 \lambda^2$ of the conformal mapping in (2.1) can be computed from those of $\nabla(\nabla \cdot \underline{v})$ and $\nabla \times \nabla \times \underline{v}$ by the use of formulas given by WHITHAM (1963). These computations lead to the following physical components, denoted by enclosing the index in parentheses.

$$\begin{aligned} [\nabla(\nabla \cdot \underline{v})](i) &= \sqrt{J} \frac{\partial f_1}{\partial \lambda^i} , \\ [\nabla \times \nabla \times \underline{v}](i) &= \sqrt{J} \left[\delta_i^1 \left(\frac{\partial f_2}{\partial \lambda^2} \right) - \delta_i^2 \left(\frac{\partial f_1}{\partial \lambda^1} \right) \right] , \end{aligned} \quad (2.18)$$

$$\begin{aligned} f_1 &= J \left[\frac{\partial}{\partial \lambda^1} (v(1)/\sqrt{J}) + \frac{\partial}{\partial \lambda^2} (v(2)/\sqrt{J}) \right] , \\ f_2 &= J \left[\frac{\partial}{\partial \lambda^1} (v(2)/\sqrt{J}) - \frac{\partial}{\partial \lambda^2} (v(1)/\sqrt{J}) \right] , \end{aligned} \quad (2.19)$$

where δ_j^i is the Kronecker delta and $i=1,2$. The quantities $v(1)$ and $v(2)$ are the physical components of \underline{v} corresponding to λ^1 and λ^2 . Thus, we obtain

$$[\nabla \cdot \underline{t}]_V(i) = \alpha_0 \sqrt{J} \left(\frac{\partial f_1}{\partial \lambda^i} \right) + \alpha_1 \sqrt{J} \left[2 \left(\frac{\partial f_1}{\partial \lambda^i} \right) - \delta_i^1 \left(\frac{\partial f_2}{\partial \lambda^2} \right) + \delta_i^2 \left(\frac{\partial f_2}{\partial \lambda^1} \right) \right], \quad (2.20)$$

where the suffix V on the left-hand side denotes the value of the physical component of $\nabla \cdot \underline{t}$ obtained by using the vector approach and $i=1,2$.

Tensor Approach

The divergence of \underline{t} in the tensor approach is obtained by using the following well known expression, see SYNGE and SCHILD (1949)

$$[\nabla \cdot \underline{t}]^\ell = t^{k\ell}_{;k} = \frac{\partial t^{k\ell}}{\partial \lambda^k} + t^{km} \{^{\ell}_{mk}\} + t^{m\ell} \{^k_{mk}\}, \quad (2.21)$$

where $[\nabla \cdot \underline{t}]^\ell$ denotes the ℓ th contravariant component of $\nabla \cdot \underline{t}$. Substituting (2.10) into (2.9) and the resulting expression into (2.21), and then applying the well known Ricci's theorem of tensor analysis [SYNGE and SCHILD (1949)],

$$g^{k\ell}_{;m} = 0, \quad (2.22)$$

for all k, ℓ , and m , one obtains

$$[\nabla \cdot \underline{t}]^\ell = \alpha_0 \sqrt{J} \frac{\partial f_1}{\partial \lambda^\ell} g^{\ell\ell} + \alpha_1 \left[\frac{\partial t^{k\ell}}{\partial \lambda^k} + I^{km} \{^{\ell}_{mk}\} + I^{m\ell} \{^k_{mk}\} \right], \quad (2.23)$$

where $I^{k\ell}$, I^{km} , and $I^{m\ell}$ are all obtainable from the general expression

$$I^{ij} = v^i_{;j} g^{jj} + v^j_{;i} g^{ii} . \quad (2.24)$$

It may readily be identified that the bracketed expression in the coefficient of α_1 in (2.23) is the covariant derivative and divergence of the tensor with respect to λ^k , that is,

$$I^{k\ell}_{;k} = \frac{\partial I^{k\ell}}{\partial \lambda^k} + I^{km}_{\{mk\}} + I^{m\ell}_{\{mk\}} . \quad (2.25)$$

For practical applications, one needs to obtain the physical components from the vector and tensor components. The latter are related to the former by [see, TRUESDELL (1953, 1954) and ERICKSEN (1960)]

$$\begin{aligned} v^k &= v(k)/\sqrt{g_{kk}} , \quad v_k = v(k)\sqrt{g_{kk}} , \\ t^{k\ell} &= t_{(k)(\ell)}/\sqrt{g_{kk}g_{\ell\ell}} , \quad t_{k\ell} = t_{(k)(\ell)}\sqrt{g_{kk}g_{\ell\ell}} , \quad t^k_{\ell} = t_{(k)(\ell)}\sqrt{\frac{g_{\ell\ell}}{g_{kk}}} , \end{aligned} \quad (2.26)$$

where the indices enclosed in parentheses, as before, represent physical components. Thus, in tensor approach, the physical components of $\nabla \cdot \underline{t}$ are obtained from (2.23) leading to

$$[\nabla \cdot \underline{t}]_T(i) = \alpha_0 \sqrt{J} \frac{\partial f_1}{\partial \lambda^i} + \alpha_1 I^{ki}_{;k} \sqrt{g_{ii}} , \quad i, k = 1, 2 , \quad (2.27)$$

where the suffix T in the left-hand side of (2.27) denotes the value of the physical component of $\nabla \cdot \underline{t}$ obtained by using the tensor approach.

Since the values of the physical components of $\nabla \cdot \underline{t}$ obtained by using the vector and tensor approaches are required to be identical, we must have from (2.20) and (2.27)

$$[\nabla \cdot \underline{t}]_T(i) - [\nabla \cdot \underline{t}]_V(i) = \alpha_1 [I^{ki}_{;k} \sqrt{g_{ii}} - \sqrt{J} (2 \frac{\partial f_1}{\partial \lambda^1} + \delta^1_i \frac{\partial f_2}{\partial \lambda^2} + \delta^2_i \frac{\partial f_2}{\partial \lambda^1})] \quad (2.28)$$

$$= 0 ,$$

for the coefficients of α_0 in the two terms on the left-hand side of (2.28) being identical automatically cancel out. The coefficient of α_1 in (2.28) can be evaluated by using (2.19) and (2.24)-(2.26). After somewhat lengthy but straightforward algebra, one readily finds that (2.28) finally reduces to

$$v^2 J - (1/J) [(\partial J / \partial \lambda^1)^2 + (\partial J / \partial \lambda^2)^2] = 0 , \quad (2.29)$$

where

$$v^2 = \frac{\partial^2}{\partial (\lambda^1)^2} + \frac{\partial^2}{\partial (\lambda^2)^2} , \quad (2.30)$$

is the familiar Laplacian operator of the plane rectangular Cartesian system (λ^1, λ^2) .

It can now be readily verified that the Jacobian of a transformation which is given to be conformal readily satisfies the condition (2.29). Conversely, by retracing the steps, one readily finds that if the Jacobian of an arbitrary plane coordinate transformation satisfies (2.29) then the latter must be a

conformal one. Thus eq. (2.29) is an invariant geometric property of the Jacobian since it is independent of the choice of the conformal map F in (2.1).

III. APPLICATIONS

In this section, two physical examples are presented, one from fluid mechanics and another from solid mechanics, in order to illustrate the above property of the Jacobian of the conformal transformations involved.

IIIa. Conformal mapping of the flow between eccentric rotating cylinders

The flow between eccentric rotating cylinders is most conveniently studied by introducing a modified bipolar coordinate system, see WANNIER (1950), WOOD (1957), and DIPRIMA and STUART (1972). Here the main interest consists only in the conformal mapping in terms of the modified bipolar coordinate system and in the derivation of the condition which must be satisfied by the Jacobian of the transformation. The momentum equations for a laminar flow in terms of the stress field are also formulated.

Consider the flow between two infinitely long circular cylinders of radii a and b ($b > a$) with centers set a distance ae apart. In order to insure that the cylinders do not touch we require $ae < b - a$, which can be written

$$0 \leq \epsilon < 1, \quad (3.1)$$

where

$$\epsilon = e/\delta, \quad \delta = (b - a)/a. \quad (3.2)$$

The parameter ϵ is the eccentricity and the parameter δ , called the clearance ratio, is a measure of the ratio of the mean clearance between the cylinders to the radius of the inner cylinder. Let the polar coordinate system have its origin at the axis of the inner cylinder with the ray $\theta = 0$ through the axis of the outer cylinder.

Following Wood in further detail we introduce modified bipolar coordinates by means of the conformal transformation

$$z = a(\zeta + \gamma)/(1 + \gamma\zeta), \quad z = x + iy = re^{i\theta}, \quad \zeta = \xi + i\eta = \rho e^{i\phi}, \quad (3.3)$$

where γ is a real constant given by

$$\gamma = (1/2\epsilon)\{-[2 + \delta(1 - \epsilon)] + [(2 + \delta(1 - \epsilon))^2 - 4\epsilon^2]^{1/2}\} \quad (3.4)$$

The coordinate curves $\rho = \text{constant}$ are circles; in particular the inner and outer cylinders are given by $\rho = 1$ and $\rho = \beta$, respectively, where

$$\beta = \frac{1 + \delta + \epsilon\delta - \gamma}{1 - (1 + \delta)\gamma - \epsilon\gamma\delta} \quad (3.5)$$

An advantage of the modified bipolar coordinate system as compared to the usual bipolar coordinate system is that in the limit $\epsilon \rightarrow 0$ the ρ, ϕ coordinate system reduces to the r, θ coordinate system except for a scale factor a .

The Jacobian J , of the transformation (3.3) is given by

$$J = (1 + 2\gamma\rho\cos\phi + \gamma^2\rho^2)^2/(1 - \gamma^2)^2, \quad |\gamma| \neq 1. \quad (3.6)$$

The element of arc length in two dimensions is

$$ds^2 = dr^2 + r^2 d\theta^2 = a^2 J^{-1} (d\rho^2 + \rho^2 d\phi^2) \quad (3.7)$$

The stress components for an incompressible viscous fluid in terms of the modified bipolar coordinate system ρ, ϕ are obtained from the linear constitutive relation having the form (2.7) and (2.8) with $\alpha_0 \text{tr} \underline{d} = -p$ and $\alpha_1 = \mu$, where \underline{t} = stress tensor, \underline{d} = deformation rate tensor, \underline{v} = velocity vector, p = fluid pressure and μ = viscosity coefficient (assumed to be independent of spatial coordinates).

In the vector approach, $\nabla \cdot \underline{t}$ takes the form (2.17). Now let v_ρ and v_ϕ be the velocity components in the ρ and ϕ directions, respectively. Then the component forms of $\nabla \cdot \underline{t}$ become, in the ρ and ϕ directions,

$$[\nabla \cdot \underline{t}]_V(\rho) = -\frac{\sqrt{J}}{a} \frac{\partial p}{\partial \rho} + \frac{\mu}{a^2} \sqrt{J} (g_1 + \frac{1}{\sqrt{J}} g_2 + a g_3) \quad (3.8)$$

where

$$\begin{aligned} g_1 &= \frac{\partial}{\partial \rho} \left[\frac{J}{\rho} \frac{\partial}{\partial \rho} \left(\frac{\rho v_\rho}{\sqrt{J}} \right) \right] + \frac{1}{\rho^2} \frac{\partial}{\partial \phi} \left[J \frac{\partial}{\partial \phi} \left(-\frac{v_\rho}{\sqrt{J}} \right) \right] , \\ g_2 &= -\frac{2J}{\rho^2} \frac{\partial v_\phi}{\partial \phi} + \frac{1}{\rho} \frac{\partial J}{\partial \rho} \frac{\partial v_\phi}{\partial \phi} - \frac{1}{\rho} \frac{\partial J}{\partial \phi} \frac{\partial v_\phi}{\partial \rho} , \\ g_3 &= \frac{\partial}{\partial \rho} \left[\frac{J}{a\rho} \left\{ \frac{\partial}{\partial \rho} \left(\frac{\rho v_\rho}{\sqrt{J}} \right) + \frac{\partial}{\partial \phi} \left(-\frac{v_\phi}{\sqrt{J}} \right) \right\} \right] , \end{aligned} \quad (3.9)$$

and

$$[\nabla \cdot \underline{t}]_V(\phi) = -\frac{\sqrt{J}}{a\rho} \frac{\partial p}{\partial \phi} + \frac{\mu\sqrt{J}}{a^2} (h_1 + \frac{1}{\sqrt{J}} h_2 + a h_3) \quad (3.10)$$

where

$$\begin{aligned}
 h_1 &= \frac{\partial}{\partial \rho} \left[\frac{J}{\rho} \frac{\partial}{\partial \rho} \left(\frac{\rho v_\phi}{\sqrt{J}} \right) \right] + \frac{1}{\rho^2} \frac{\partial}{\partial \phi} \left[J \frac{\partial}{\partial \phi} \left(\frac{v_\phi}{\sqrt{J}} \right) \right], \\
 h_2 &= \frac{2J}{\rho^2} \frac{\partial v_\rho}{\partial \phi} - \frac{1}{\rho} \frac{\partial J}{\partial \rho} \frac{\partial v_\rho}{\partial \phi} + \frac{1}{\rho} \frac{\partial J}{\partial \phi} \frac{\partial v_\rho}{\partial \rho}, \\
 h_3 &= \frac{1}{\rho} \frac{\partial}{\partial \phi} \left[\frac{J}{\rho} \left\{ \frac{\partial}{\partial \rho} \left(\frac{\rho v_\rho}{\sqrt{J}} \right) + \frac{\partial}{\partial \phi} \left(\frac{v_\phi}{\sqrt{J}} \right) \right\} \right],
 \end{aligned} \tag{3.11}$$

In the tensor approach, the stress constitutive relation takes the form (2.9) and (2.10) with $\alpha_0 \text{trd} = -p$, $\alpha_1 = \mu$. The expressions for $[\nabla \cdot \underline{t}]_T(\rho)$ and $[\nabla \cdot \underline{t}]_T(\phi)$ are obtained from (2.27). Thus,

$$[\nabla \cdot \underline{t}]_T(\rho) = -\frac{\sqrt{J}}{a} \frac{\partial p}{\partial \rho} + \frac{\mu \sqrt{J}}{a^2} (\tau_1 + \tau_2), \tag{3.12}$$

where

$$\begin{aligned}
 \tau_1 &= 2 \frac{\partial}{\partial \rho} \left(\sqrt{J} \frac{\partial v_\rho}{\partial \rho} \right) + \frac{2\sqrt{J}}{\rho} \frac{\partial v_\rho}{\partial \rho} - \frac{2v_\rho}{\rho^2} \sqrt{J} + \frac{2v_\rho}{\rho \sqrt{J}} \frac{\partial J}{\partial \rho} \\
 &\quad - \frac{1}{\sqrt{J}} \frac{\partial J}{\partial \rho} \frac{\partial v_\rho}{\partial \rho} - \frac{v_\rho}{2J\sqrt{J}} \left(\frac{\partial J}{\partial \rho} \right)^2,
 \end{aligned} \tag{3.13}$$

$$\begin{aligned}
 \tau_2 &= -\frac{2}{\rho \sqrt{J}} \frac{\partial J}{\partial \phi} \frac{\partial v_\phi}{\partial \rho} + \frac{v_\phi}{2\rho^2 \sqrt{J}} \frac{\partial J}{\partial \phi} + \frac{v_\phi}{4\rho J \sqrt{J}} \frac{\partial J}{\partial \rho} \frac{\partial J}{\partial \phi} \\
 &\quad - \frac{v_\phi}{2\rho \sqrt{J}} \frac{\partial^2 J}{\partial \rho \partial \phi} - \frac{3\sqrt{J}}{\rho^2} \frac{\partial v_\phi}{\partial \phi} + \frac{3}{2\rho \sqrt{J}} \frac{\partial v_\phi}{\partial \phi} \frac{\partial J}{\partial \rho} \\
 &\quad + \frac{\sqrt{J}}{\rho} \frac{\partial^2 v_\phi}{\partial \phi^2} + \frac{1}{\rho} \frac{\partial}{\partial \phi} \left(\sqrt{J} \frac{\partial v_\phi}{\partial \rho} \right) - \frac{3}{4\rho^2} \frac{\partial v_\phi}{J \sqrt{J}} \left(\frac{\partial J}{\partial \phi} \right)^2 \\
 &\quad + \frac{v_\rho}{2\rho^2 \sqrt{J}} \frac{\partial^2 J}{\partial \phi^2},
 \end{aligned} \tag{3.14}$$

and a corresponding expression for $[\nabla \cdot \underline{t}]_T(\phi)$ which is obtained in like manner. Now, since the vector and tensor approaches must produce identical results for $\nabla \cdot \underline{t}$, the difference between (3.8) and (3.12) must vanish as also the difference between (3.10) and the corresponding expression for $[\nabla \cdot \underline{t}]_T(\phi)$ must vanish for all possible arbitrary velocity fields. Thus

$$[\nabla \cdot \underline{t}]_T(\rho) - [\nabla \cdot \underline{t}]_V(\rho) = 0, \quad (3.15)$$

yields the condition

$$\frac{\partial^2 J}{\partial \rho^2} + \frac{1}{\rho} \cdot \frac{\partial J}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 J}{\partial \phi^2} - \frac{1}{J} \left[\left(\frac{\partial J}{\partial \rho} \right)^2 + \frac{1}{\rho^2} \left(\frac{\partial J}{\partial \phi} \right)^2 \right] = 0. \quad (3.16)$$

Similarly, it can be readily verified that

$$[\nabla \cdot \underline{t}]_T(\phi) - [\nabla \cdot \underline{t}]_V(\phi) = 0, \quad (3.17)$$

yields the same identical condition (3.16). This establishes the constraint on the Jacobian of the conformal mapping given in (3.3).

For purposes of future applications, it is useful to formulate here the momentum equations in terms of the stress divergence components employing the modified bipolar coordinates ρ, ϕ :

$$\begin{aligned} \frac{\partial v_\rho}{\partial t} + \frac{\sqrt{J}}{a} v_\rho \frac{\partial v_\rho}{\partial \rho} + \frac{\sqrt{J}}{a\rho} v_\phi \frac{\partial v_\rho}{\partial \phi} - \frac{1}{a} \frac{\partial}{\partial \phi} \left(\frac{\sqrt{J}}{\rho} \right) v_\phi v_\rho + \frac{\rho}{a} \frac{\partial}{\partial \rho} \left(\frac{\sqrt{J}}{\rho} \right) v_\phi^2 \\ = \frac{\sqrt{J}}{a\sigma} \left[\frac{\partial t_{\rho\rho}}{\partial \rho} + \left(\frac{1}{\rho} - \frac{1}{2J} \frac{\partial J}{\partial \rho} \right) (t_{\rho\rho} - t_{\phi\phi}) + \frac{1}{\rho} \left(\frac{\partial t_{\rho\phi}}{\partial \phi} - \frac{t_{\rho\phi}}{J} \frac{\partial J}{\partial \phi} \right) \right], \end{aligned} \quad (3.18)$$

$$\begin{aligned}
& \frac{\partial v}{\partial t} \phi + \frac{\sqrt{J}}{a} v_{\rho} \frac{\partial v}{\partial \rho} \phi + \frac{\sqrt{J}}{a \rho} v_{\phi} \frac{\partial v}{\partial \phi} \phi - \frac{\rho}{a} \frac{\partial}{\partial \rho} \left(\frac{\sqrt{J}}{\rho} \right) v_{\rho} v_{\phi} + \frac{1}{a} \frac{\partial}{\partial \phi} \left(\frac{\sqrt{J}}{\rho} \right) v_{\rho}^2 \\
& = \frac{\sqrt{J}}{a v} \left\{ \frac{1}{\rho} \frac{\partial t}{\partial \phi} \phi - \frac{\partial t}{\partial \rho} \phi \right\} + \left(\frac{2}{\rho} - \frac{1}{J} \frac{\partial J}{\partial \rho} \right) t_{\rho \phi} + \frac{1}{2 \rho J} \frac{\partial J}{\partial \phi} (t_{\rho \rho} - t_{\phi \phi})] \quad , \quad (3.19)
\end{aligned}$$

where t denotes the time variable and $\sigma =$ mass density of the fluid and

$t_{\rho \rho}$, $t_{\rho \phi}$ and $t_{\phi \phi}$ are the stress components given by

$$t_{\rho \rho} = -p + \frac{2\mu}{a} \left[\sqrt{J} \frac{\partial v}{\partial \rho} - v_{\phi} \frac{\partial}{\partial \phi} \left(\frac{\sqrt{J}}{\rho} \right) \right] ,$$

$$t_{\rho \phi} = \frac{\mu}{a} \left[\sqrt{J} \frac{\partial v}{\partial \phi} + \rho v_{\phi} \frac{\partial}{\partial \rho} \left(\frac{\sqrt{J}}{\rho} \right) + \sqrt{J} \frac{\partial v}{\partial \rho} \phi + v_{\rho} \frac{\partial}{\partial \phi} \left(\frac{\sqrt{J}}{\rho} \right) \right] , \quad (3.20)$$

$$t_{\phi \phi} = -p + \frac{2\mu}{a} \left[\sqrt{J} \frac{\partial v}{\partial \phi} - \rho v_{\rho} \frac{\partial}{\partial \rho} \left(\frac{\sqrt{J}}{\rho} \right) \right] .$$

The equation of continuity is given by

$$\frac{\partial}{\partial \rho} \left(\frac{\rho}{\sqrt{J}} v_{\rho} \right) + \frac{\partial}{\partial \phi} \left(\frac{v_{\phi}}{\sqrt{J}} \right) = 0 . \quad (3.21)$$

IIIb. Conformal mapping for the torsion of a hollow eccentric shaft

Consider a prism in the form of a hollow shaft with the inner and outer surfaces being long cylinders which are not concentric. Suppose that the

material of the cylinders is made up of a linear homogeneous isotropic elastic solid characterized by Hooke's law with constant lame coefficients. The prism is twisted by couples applied at the ends of the prism. It is well known, see TIMOSHENKO and GOODIER (1970), that the deformation of the twisted shaft consists (1) of rotations of cross sections of the shaft and (2) of warping of the cross sections assumed to be the same for all cross sections. In order to delineate the warping of the cross sections, a function known as the torsion function is introduced. The determination of this function is an integral part of the problem. The solution of this problem is most conveniently effected by introducing a conformal transformation in the plane of a cross section of the prism

$$z = c \cdot \tan \frac{1}{2} \zeta, \quad z = x + iy, \quad \zeta = \xi + i\eta, \quad (3.22)$$

where $2c$ is the distance between the poles of the bipolar coordinate system ξ, η induced by (3.22). In the present study, we are only interested in the property of the Jacobian of the above conformal transformation.

The Jacobian of the transformation is readily found using (2.2).

$$J = \frac{1}{c^2} (\cos \xi + \cosh \eta)^2 > 0. \quad (3.23)$$

The curves $\eta = \text{constant}$ are found to be coaxial circles with equations

$$x^2 + (y - c \cdot \coth \eta)^2 = c^2 \operatorname{cosech}^2 \eta, \quad (3.24)$$

while the curves $\xi = \text{constant}$ are circles given by

$$(x + c \cdot \cot \xi)^2 + y^2 = c^2 \operatorname{cosec}^2 \xi, \quad (3.25)$$

and the cross section of the prism is bounded, say, by the eccentric circles

$$\eta = \eta_0 \text{ and } \eta = \eta_1 (<\eta_0) .$$

The stress-strain relations and other pertinent formulas of linear elasticity are all analogous to the formulas (2.15)-(2.17) where \underline{t} is the stress tensor, \underline{d} is replaced by the strain tensor \underline{e} , the vector \underline{y} is replaced by the displacement vector \underline{u} and α_0 and α_1 are respectively, replaced by the Lamé coefficients λ_e and μ_e .

Following the general method given in section 2 one obtains the expressions for stress divergence using the vector and tensor approaches. Thus

$$\begin{aligned} [\underline{V} \cdot \underline{t}]_V(\xi) &= \sqrt{J} [(\lambda_e + 2\mu_e) \frac{\partial}{\partial \xi} \{ J \frac{\partial}{\partial \xi} (\frac{u_\xi}{\sqrt{J}}) + J \frac{\partial}{\partial \eta} (\frac{u_\eta}{\sqrt{J}}) \} \\ &\quad - \mu_e \frac{\partial}{\partial \eta} \{ J \frac{\partial}{\partial \xi} (\frac{u_\eta}{\sqrt{J}}) - J \frac{\partial}{\partial \eta} (\frac{u_\xi}{\sqrt{J}}) \}] , \end{aligned} \quad (3.26)$$

$$\begin{aligned} [\underline{V} \cdot \underline{t}]_V(\eta) &= \sqrt{J} [(\lambda_e + 2\mu_e) \frac{\partial}{\partial \eta} \{ J \frac{\partial}{\partial \xi} (\frac{u_\xi}{\sqrt{J}}) + J \frac{\partial}{\partial \eta} (\frac{u_\eta}{\sqrt{J}}) \} \\ &\quad + \mu_e \frac{\partial}{\partial \xi} \{ J \frac{\partial}{\partial \xi} (\frac{u_\eta}{\sqrt{J}}) - J \frac{\partial}{\partial \eta} (\frac{u_\xi}{\sqrt{J}}) \}] , \end{aligned} \quad (3.27)$$

where u_ξ and u_η are the components of the displacement vector relative to the bipolar coordinate system ξ and η .

In the tensor approach,

$$[\nabla \cdot \underline{t}]_T(\xi) = \sqrt{J} \left[\frac{\partial t_{\xi\xi}}{\partial \xi} - \frac{1}{2J} \frac{\partial J}{\partial \xi} (t_{\xi\xi} - t_{\eta\eta}) + \frac{\partial t_{\xi\eta}}{\partial \eta} - \frac{1}{J} \frac{\partial J}{\partial \eta} t_{\xi\eta} \right], \quad (3.28)$$

$$[\nabla \cdot \underline{t}]_T(\eta) = \sqrt{J} \left[\frac{\partial t_{\eta\eta}}{\partial \eta} + \frac{1}{2J} \frac{\partial J}{\partial \eta} (t_{\xi\xi} - t_{\eta\eta}) + \frac{\partial t_{\xi\eta}}{\partial \xi} - \frac{1}{J} \frac{\partial J}{\partial \xi} t_{\xi\eta} \right]. \quad (3.29)$$

$$t_{\xi\xi} = \sqrt{J} \left[\lambda_e \left(\frac{\partial u_\xi}{\partial \xi} + \frac{\partial u_\eta}{\partial \eta} - \frac{1}{2J} u_\xi \frac{\partial J}{\partial \xi} - \frac{1}{2J} u_\eta \frac{\partial J}{\partial \eta} \right) + 2\mu_e \left(\frac{\partial u_\xi}{\partial \xi} - \frac{1}{2J} u_\eta \frac{\partial J}{\partial \eta} \right) \right], \quad (3.30)$$

$$t_{\xi\eta} = \sqrt{J} \mu_e \left(\frac{\partial u_\xi}{\partial \eta} + \frac{1}{2J} \frac{\partial J}{\partial \xi} u_\eta + \frac{\partial u_\eta}{\partial \xi} + \frac{1}{2J} u_\xi \frac{\partial J}{\partial \eta} \right), \quad (3.31)$$

$$t_{\eta\eta} = \sqrt{J} \left[\lambda_e \left(\frac{\partial u_\xi}{\partial \xi} + \frac{\partial u_\eta}{\partial \eta} - \frac{1}{2J} u_\xi \frac{\partial J}{\partial \xi} - \frac{1}{2J} u_\eta \frac{\partial J}{\partial \eta} \right) + 2\mu_e \left(\frac{\partial u_\eta}{\partial \eta} - \frac{1}{2J} u_\xi \frac{\partial J}{\partial \xi} \right) \right]. \quad (3.32)$$

Substituting (3.30) - (3.32) into (3.28), one obtains

$$[\nabla \cdot \underline{t}]_T(\xi) = \lambda_e E_1 + \mu_e E_2, \quad (3.33)$$

where

$$\begin{aligned}
 E_1 = & J \frac{\partial^2 u_\xi}{\partial \xi^2} + J \frac{\partial^2 u_\eta}{\partial \xi \partial \eta} - \frac{1}{2} u_\xi \frac{\partial^2 J}{\partial \xi^2} + \frac{1}{4J} \left(\frac{\partial J}{\partial \xi} \right)^2 u_\xi \\
 & - \frac{1}{2} u_\eta \frac{\partial^2 J}{\partial \xi \partial \eta} - \frac{1}{2} \frac{\partial J}{\partial \eta} \frac{\partial u_\eta}{\partial \xi} + \frac{1}{4J} \frac{\partial J}{\partial \xi} \frac{\partial J}{\partial \eta} u_\eta \\
 & + \frac{1}{2} \frac{\partial J}{\partial \xi} \frac{\partial u_\eta}{\partial \eta} , \tag{3.34}
 \end{aligned}$$

$$\begin{aligned}
 E_2 = & 2J \frac{\partial^2 u_\xi}{\partial \xi^2} - \frac{3}{2} \frac{\partial J}{\partial \eta} \frac{\partial u_\eta}{\partial \xi} - \frac{1}{2} u_\eta \frac{\partial^2 J}{\partial \xi \partial \eta} + \frac{1}{4J} \frac{\partial J}{\partial \xi} \frac{\partial J}{\partial \eta} u_\eta \\
 & + \frac{3}{2} \frac{\partial J}{\partial \xi} \frac{\partial u_\eta}{\partial \eta} - \frac{1}{2J} u_\xi \left(\frac{\partial J}{\partial \xi} \right)^2 - \frac{3}{4J} u_\xi \left(\frac{\partial J}{\partial \eta} \right)^2 \\
 & + J \frac{\partial^2 u_\xi}{\partial \eta^2} + J \frac{\partial^2 u_\eta}{\partial \xi \partial \eta} + \frac{1}{2} u_\xi \frac{\partial^2 J}{\partial \eta^2} . \tag{3.35}
 \end{aligned}$$

Subtracting (3.26) from (3.33), one obtains

$$[\nabla \cdot \underline{t}]_T(\xi) - [\nabla \cdot \underline{t}]_V(\xi) = 0 , \tag{3.36}$$

which yields for all possible arbitrary choices of u_ξ

$$\nabla^2 J - \frac{1}{J} \left[\left(\frac{\partial J}{\partial \xi} \right)^2 + \left(\frac{\partial J}{\partial \eta} \right)^2 \right] = 0 . \tag{3.37}$$

In like manner, the difference expression,

$$[\nabla \cdot \underline{t}]_T(\eta) - [\nabla \cdot \underline{t}]_V(\eta) = 0 , \tag{3.38}$$

also leads to the same identical condition as in (3.37) on the Jacobian.

IV. DISCUSSION

From the foregoing analysis, the following observations and conclusions can be effected regarding the properties of the Jacobian of the conformal transformations.

(1) The condition satisfied by the Jacobian given by (2.29) is entirely independent of the mapping function of the conformal transformation and hence is invariant with respect to arbitrary choices of conformal maps.

(2) Also, the condition on J is independent of all physical fields and parameters within the class of continuously differentiable, symmetric tensor-valued functions.

Hence the constraint on the Jacobian is an invariant geometric condition which arises purely as a geometric property of the conformal transformation and its Jacobian.

(3) Only the coefficient of α_0 in (2.28) vanishes identically without constraining J in any manner, but the coefficient of α_1 involves the condition on J . The phenomenon can be accounted for by observing that the coefficient of α_0 involves the first invariant, $\text{tr} \underline{d}$, and owing to this invariance, there occurs no change in the geometry of the region. On the other hand, the coefficient of α_1 is indeed associated with a geometry change since it leads to a constraint on the Jacobian of the transformation causing such a change. If for example, \underline{v} is the velocity vector of a fluid particle in a given region of flow, the coefficient $\nabla(\nabla \cdot \underline{v})$ of α_0 does not involve any distortional

change since it represents a dilatation rate. On the other hand, the coefficient of α_1 does represent a change of shape which means a change in geometry implying a constraining condition on the Jacobian of the transformation causing the change of shape.

Thus, the vector and tensor approaches yield identical results for the divergence of a second rank tensor \underline{t} of the type described earlier if and only if the invariant geometric condition (2.29) on the Jacobian of the conformal mapping in (2.1) is satisfied.

ACKNOWLEDGEMENT

The authors wish to thank the U.S. Army Research Office (ARO) for their support. One of us (MNLN) is spending a year as an IPA at the ARO.

REFERENCES

1. A. E. H. LOVE, A Treatise on the Mathematical Theory of Elasticity, Torsion, Chapter 14, Dover Publications, Fourth Edition, New York, 1944.
2. N. I. MUSKHELISHVILI, Some Fundamental Problems of the Theory of Elasticity, English Translation, North Holland Publishing Company, 1954.
3. J. L. SYNGE and A. SCHILD, Tensor Calculus, University of Toronto Press, 1949.
4. C. A. TRUESDELL, Physical Components of Vectors and Tensors, ZAMM, Vol. 33 (1953), 345-356.
5. Ibid - ZAMM, Vol. 34 (1954), 69-70.
6. J. L. ERICKSEN, Tensor Fields, Appendix in the Classical Field Theories by C. A. TRUESDELL and R. TOUPIN, Handbuch der Physik, Vol. 3, Part 1, (1960).

7. G. B. WHITHAM, The Navier-Stokes Equations of Motion, Chapter 3 of Laminar Boundary Layers, L. ROSENHEAD, Editor, Clarendon Press, Oxford (1963).
8. G. H. WANNIER, A Contribution to the Hydrodynamics of Lubrication, Quarterly of Applied Mathematics, Vol. 8 (1950), 1-32.
9. W. W. WOOD, The Asymptotic Expansions at Large Reynolds Numbers for Steady Motion Between Non-Coaxial Rotating Cylinders, Journal of Fluid Mechanics, Vol. 3 (1957), 159-175.
10. R. C. DIPRIMA and J. T. STUART, Flow Between Eccentric Rotating Cylinders, Journal of Lubrication Technology, Transactions of the ASME, (1972), 266-274.
11. S. P. TIMOSHENKO and GOODIER, Theory of Elasticity, Torsion, Chapter 10, McGraw Hill Book Company, New York (1970).

LARGE ELASTIC DEFORMATION OF A SHEET DUE TO FLUID LOAD

EDWARD W. ROSS, JR.
Mathematician
US Army Natick Research & Development Center
Natick, Massachusetts 01760-5017

Abstract

This paper analyzes the behavior of a tarpaulin, suspended at its ends when a puddle of water accumulates on it during a rainstorm. The tarpaulin is taken as a wide, thin sheet of linearly elastic material, without bending strength, initially horizontal and unstressed. It undergoes large, plane deformation, caused by the weight of the puddle, and is idealized as a string. The solution is found in closed form except for one boundary condition that has to be satisfied by trial and error and involving a numerical integration. The deformed sheet has the shape of a sine function beneath the load. Asymptotic formulas are derived in the small deformation limit, but most of these are sufficiently accurate to be useful for practical ranges of the parameters when the deformations are large.

1. Introduction

Tents, tarpaulins and other forms of thin, flexible sheets are much used by the military for the shelter of troops and equipment. The theoretical behavior of such structures under conditions commonly met in the field, such as rain, snow and wind, has not been studied extensively. The present paper is an investigation of one such situation, namely, when rain falls on a wide flat, horizontal sheet, fixed at the edges. Under these conditions the sheet sags and water accumulates in a puddle near the center. We analyze a simple version of this process, i.e. one-dimensional continuum (a string) under a liquid, gravity load that varies along the string in a manner consistent with the condition that the liquid surface be horizontal. The deformations are not required to be small, but a linearly elastic constitutive relation is assumed. Bending and dynamic effects are omitted.

The behavior of a flexible string has been analyzed by many writers, going back to the times of Bernoulli and Love. Various aspects of the problem have been studied recently by Antman [1] and Pugsley [2] among others. Much effort has gone into the development of finite-element programs for the numerical solution of nonlinear string problems, see, for example, Fried [3], Huddleston [4] and Peyrot and Goulois [5]. The author is aware of only one paper dealing with a problem similar to the present one, Malcolm and Glockner [6], and that is restricted to small deformation.

2. Formulation

A wide sheet is idealized as a one-dimensional continuum, a string, of length $2L$, initially straight and horizontal, with its left end at $x = y = 0$ and mid-point at $x = L, y = 0$, see Figure 1. The sheet is deformed symmetrically about the mid-point by a puddle of liquid with a horizontal surface and variable depth. Variables X, Y are the deformed coordinates of the point on the string that is initially at (x, y) .

The kinematics of the deformed string are specified in terms of θ , where

$$\tan \theta = dY/dX = \text{slope of string} \quad (1)$$

If dS is the deformed length of a small string element whose initial length was dx , then

$$dS/dx = 1 + \epsilon = [(dX/dx)^2 + (dY/dx)^2]^{1/2} \quad (2)$$

where ϵ is the strain. Also

$$dX/dx = (1 + \epsilon) \cos \theta, \quad dY/dx = (1 + \epsilon) \sin \theta \quad (3)$$

We do not assume that $|\epsilon| \ll 1$.

The constitutive law for linear elasticity is used,

$$T = \text{tension in string} = T_0 \epsilon \quad (4)$$

although it is conceivable that some progress could be made with a more general elastic law.

Equilibrium of a string element in the X and Y directions implies

$$\begin{aligned} dT/dx &= -F_y (dS/dx) \sin \theta \\ T d\theta/dx &= -F_y (dS/dx) \cos \theta \end{aligned} \quad (5)$$

where $F_y dS$ is the vertical upward force of the liquid on the deformed string-element of length dS . We assume that the water surface is a distance below the initial position of the string (see Figure 2), and the wet portion of the string has initial coordinates

$$AL \leq x \leq L.$$

If μ is the density of the liquid, and U is the unit step function, we find

with the aid of (3)

$$\begin{aligned} F_y dS/dx &= \mu(Y + \Delta) (dX/dx) U(x - AL) \\ &= \mu(Y + \Delta) (1 + \epsilon) \cos \theta U(x - AL) \end{aligned} \quad (6)$$

The system of equations is then

$$\begin{aligned} dT/dx &= -\mu(Y + \Delta) (1 + \epsilon) \sin \theta \cos \theta U(x - AL) \\ T d\theta/dx &= -\mu(Y + \Delta) (1 + \epsilon) \cos^2 \theta U(x - AL) \\ dY/dx &= (1 + \epsilon) \sin \theta \\ dX/dx &= (1 + \epsilon) \cos \theta \end{aligned} \quad (7)$$

These have to be solved with the boundary conditions

$$\begin{aligned} X = Y = 0 \text{ at } x = 0 \\ \theta = 0, X = L \text{ at } x = L \end{aligned} \quad (8)$$

Notice that

$$Y = -\Delta \quad \text{at } x = AL$$

because of the definitions.

The equations may be put in dimensionless form by means of the transformations

$$\begin{aligned} z = x/L, \xi = X/L, \eta = Y/L, \delta = \Delta/L, \epsilon = T/T_0 \\ \rho = L(\mu/T_0)^{1/2} \end{aligned} \quad (9)$$

after which they may be written

$$d\epsilon/dz = -\rho^2 (\eta + \delta) (1 + \epsilon) \sin \theta \cos \theta U(z - A) \quad (10)$$

$$d\theta/dz = -\rho^2 \epsilon^{-1} (\eta + \delta) (1 + \epsilon) \cos^2 \theta U(z - A) \quad (11)$$

$$d\eta/dz = (1 + \epsilon) \sin \theta \quad (12)$$

$$d\xi/dz = (1 + \epsilon) \cos \theta \quad (13)$$

with boundary conditions

$$\xi = \eta = 0 \text{ at } z = 0 \quad (14)$$

$$\theta = 0, \xi = 1 \text{ at } z = 1 \quad (15)$$

and the intermediate condition

$$\eta = -\delta \quad \text{at } z = A \quad (16)$$

The forcing function in this problem is the total load of liquid, W , in the region $0 \leq z \leq 1$, which is given by

$$W = -T \sin \theta \Big|_{z=0}$$

or, in non-dimensional terms

$$W/T_0 \equiv \lambda = -\epsilon(0) \sin \theta(0) \quad (17)$$

3. Solution

An almost-complete solution to the problem in closed form is obtainable by elementary means. The solution in the dry (i.e. unwetted) region $0 \leq z \leq A$ can be written by inspection, and the solution in $A \leq z \leq 1$ which joins smoothly with that in $0 \leq z \leq A$ can be found as a function of θ (which is monotone in $A \leq z \leq 1$) by dividing pairs of the equations.

In the dry region the solution is easily seen to be

$$\begin{aligned} \epsilon &= \epsilon(0) = \epsilon_0, \quad \theta = \theta_0 = \theta(0) \\ \xi &= (1+\epsilon_0)z \cos \theta_0, \quad \eta = (1+\epsilon_0)z \sin \theta_0 \end{aligned} \quad (18)$$

Then at $z = A$ we have

$$\begin{aligned} \epsilon &= \epsilon_0, \quad \theta = \theta_0, \quad \xi = (1+\epsilon_0)A \cos \theta_0 \\ \eta &= -\delta = (1+\epsilon_0)A \sin \theta_0 \end{aligned} \quad (19)$$

The solution for $z \geq A$ is found as follows. Dividing Equation (10) by (11), we get

$$\epsilon^{-1} d\epsilon/d\theta = \tan \theta$$

which has the solution

$$\epsilon = \epsilon_0 \cos \theta_0 \sec \theta \quad (20)$$

This shows that $\epsilon \leq \epsilon_0$, as expected. Next, if we divide (11) by (12) we see that

$$\epsilon d\theta/d\eta = -\rho^2(\eta+\delta)\cot\theta\cos\theta$$

using (20), this can be rewritten as

$$\sec^2\theta\tan\theta + Q^{-1}\rho^2(\eta+\delta)d\eta/d\theta = 0$$

and has the solution

$$\eta = -\delta - Q^{1/2}\rho^{-1}H(\theta, \theta_0) \quad (21)$$

where

$$Q = \epsilon_0 \cos\theta_0 \quad (22)$$

$$H = -\tan\theta_0(1 - \tan^2\theta/\tan^2\theta_0)^{1/2} \geq 0 \quad (23)$$

If (13) is divided by (11), the resulting equation can be written

$$d\xi/d\theta = -Q^{1/2}\rho^{-1}H^{-1}\sec^2\theta$$

which has the solution

$$\xi = A(Q + \cos\theta_0) + Q^{1/2}\rho^{-1}\psi(\theta, \theta_0) \quad (24)$$

$$\psi(\theta, \theta_0) = \pi/2 - \sin^{-1}(\tan\theta/\tan\theta_0) \quad (25)$$

At this point we have obtained solutions for ϵ, η and ξ as functions of θ which satisfy the junction conditions at $z = A$. It still remains to find the relation between θ and z . To do this, we must solve (11), which leads to

$$z = A + Q^{1/2}\rho^{-1} \int_{\theta_0}^{\theta} (Q + \cos t)^{-1} [H(t)]^{-1} \sec^2 t dt \quad (26)$$

Integrating by parts and using $\psi(\theta_0, \theta_0) = 0$, we obtain

$$z = A + Q^{1/2}\rho^{-1} [(Q + \cos\theta)^{-1} \psi(\theta, \theta_0) - \int_{\theta_0}^{\theta} (Q + \cos t)^{-2} \psi(t, \theta_0) \sin t dt] \quad (27)$$

The integrals in Equations (26) and (27) cannot be evaluated exactly but may be found by numerical integration. In doing so, the form (26) is inconvenient because of the singularity in the integrand arising from the zero of H at $\theta = \theta_0$.

The singularity is merely in the derivative of the integrand of (27) and that integral was evaluated successfully by the IMSL subroutine DCADRE.

The deformed shape of the string can be obtained in closed form by eliminating θ between (21) and (24), which leads to the relation

$$\eta = -\delta + Q^{1/2} \rho^{-1} \tan \theta_o \sin[\rho Q^{-1/2} \{\xi - A(Q + \cos \theta_o)\}] \quad (28)$$

Thus the wetted part of the sheet is shaped like a portion of a sine curve.

The parameters θ_o and ϵ_o (or, equivalently, Q) have yet to be evaluated. For this purpose we use the boundary conditions (15) in the form $\xi = z = 1$ at $\theta = 0$. These lead to the equations

$$A(Q + \cos \theta_o) + Q^{1/2} \rho^{-1} \pi/2 - 1 = 0 \quad (29)$$

$$A + Q^{1/2} \rho^{-1} [(1 + Q)^{-1} \pi/2 - I_x] = 1 \quad (30)$$

where

$$I_x = \int_{t=\theta_o}^0 (Q + \cos t)^{-2} \psi(t, \theta_o) \sin t dt \quad (31)$$

Taking A and ρ as given quantities, these equations were solved for θ_o using the IMSL version of the Brent algorithm. If θ_o is assumed, Equation (29) is merely quadratic in Q and can be solved easily. Then the function on the left of (30) is evaluated for these values of θ_o and Q , and the Brent algorithm generates successive values of θ_o until Equation (30) is satisfied with sufficient accuracy.

When the solution for θ_o has been obtained, the quantities

$$\epsilon_o = Q \sec \theta_o$$

$$\delta = -(1 + \epsilon_o) A \sin \theta_o$$

$$\lambda = -\epsilon_o \sin \theta_o$$

$$\eta(1) = -\delta + Q^{1/2} \rho^{-1} \tan \theta_o$$

can be found. The relation between z and θ is expressed by Equation (27), and Equations (21) and (24) then give the deformed positions of points parametrically in terms of their initial locations.

In applying these formulas we shall usually regard λ as the load-intensity and ρ as a parameter characteristic of the material-geometrical situation. The forms of the solutions are such that for most variables of interest the dependence on λ and ρ can only be obtained implicitly. However, it is informative to examine the small-deflection limit, for which explicit formulas can be found. This derivation requires some care and is given in the Appendix. The results are as $\lambda \rightarrow 0$

$$1 - A \sim 2^{-1/6} [\pi/(2\rho)] \lambda^{1/3} \approx 1.399 \rho^{-1} \lambda^{1/3} \quad (32)$$

$$\delta \sim -\theta_0 \sim -\eta(1) \sim (2\lambda)^{1/3} \approx 1.260 \lambda^{1/3} \quad (33)$$

$$\epsilon_0 \sim Q \sim 2^{-1/3} \lambda^{2/3} \approx 0.794 \lambda^{2/3} \quad (34)$$

4. Results and Discussion

Although the relation between initial and deformed positions of particles on the sheet is of some interest, we are primarily concerned with finding the effects of the parameters λ and ρ on the quantities A (which specifies the half-length of the puddle), θ_0 (maximum slope angle), ϵ_0 (maximum strain or dimensionless stress), $\eta(1)$ (dimensionless central deflection) and δ (dimensionless displacement of the puddle surface). Typical values for the physical parameters in anticipated applications of these results are

$$L = 3 \text{ meters}, \quad \mu = 1,000 \text{ Kg/meter}^3, \quad T_0 = 360 \text{ newton/mm}$$

Since all calculations are done per unit sheet-width perpendicular to the plane of motion, a typical value for ρ is

$$\rho = L(\mu/T_0)^{1/2} \approx 1/2.$$

First, for illustrative purposes we display in Figure 3 two examples of the deformed shape (when $\rho = 1/2$) of the sheet, for $A = 0.4$, $\lambda = .01256$, and $A = 0.8$, $\lambda = .387 \times 10^{-3}$. The actual weights of water on these sheets are 452 and 13.9 kilograms per meter of width, respectively. The more highly stressed of the two cases ($A = 0.4$) suffers a maximum strain $\epsilon_0 = 0.039$. It is thought that this sheet-material tears at about $\epsilon = 0.1$, consequently failure is not expected in either case. Comparing the initial and deformed positions of points in these graphs, we observe that the x-displacements are much smaller than the y-displacements.

Figures 4 through 8 show plots of $\epsilon_0, \theta_0, \eta(1)$, A and δ as functions of λ for $\rho = .2, .5$ and $.8$. These graphs imply two main conclusions. The first,

perhaps rather simple, is that for any ρ there is a maximum total liquid weight that the deformed sheet can carry. For, when the entire sheet is wetted, $A = \delta = 0$, any additional liquid will simply overflow at the ends of the sheet. These maxima can be seen in Figure 7 as the values of λ for which $A = 0$. The relationship between this maximum, or fully-wetted, load, λ_f , and ρ is exhibited in Figure 9. This shows that λ_f increases with ρ and, for example, $\lambda \leq \lambda_f \leq 0.3$ when $\rho \leq 0.8$.

The second main point, visible in Figures 4, 5, and 6, is that ρ does not have much effect on the relations between λ and ϵ_0, θ_0 and $n(1)$ in the range $.2 \leq \rho \leq 0.8$. This agrees with the asymptotic behaviors of ϵ_0, θ_0 and $n(1)$ for small deflections, Equations (33) and (34), which show no effect of ρ . Observe, however, that A and δ are affected by ρ . Indeed, Equation (32) and Figure 7 indicate that ρ affects the relation between A and λ for all relevant λ values, i.e. for $0 \leq \lambda \leq \lambda_f$. Equation (33) asserts that the relation between δ and λ is not influenced by ρ when $\lambda \rightarrow 0$, but the effect of ρ is felt when $\lambda \geq 1 \times 10^{-4}$, according to Figure 8.

Systematic comparisons between the asymptotic formulas (32), (33) and (34) and Figures 4 through 8 show that the asymptotic formulas for ϵ_0, θ_0 , $n(1)$ and A are almost good enough to be used throughout the range $\rho \leq 0.8$. The asymptotic formulas are of course most accurate for small λ , but the graphs show that their errors are not too bad even when λ is fairly large, say $\lambda = O(10^{-1})$. In particular, the estimate

$$\epsilon_0 = .794\lambda^{2/3}$$

is quite good even near $\lambda = 1$ and can almost be regarded as a universal formula for $\rho \leq 0.8$. The asymptotic predictions of θ_0 , $\eta(1)$ and A are less accurate but still usable as rough estimates for $\lambda = O(10^{-1})$. The simple formula (33) for δ is accurate only for very small λ and completely misses the dependence of δ on ρ .

We present next an elementary, somewhat contrived, application of these results, namely to estimate the rain conditions under which a sheet with $T = 36$ newton/cm, width = 6 meters, i.e. $L = 3$ meters, will break. We assume that the sheet tears when $\epsilon = 0.1$. The asymptotic formulas tell us that tearing occurs when

$$\lambda = (.1/.794)^{3/2} = .0447$$

$$w = \lambda T_0 = 1.609 \text{ kg/m}$$

Hence the volume of water per meter of sheet width is 1.609 m^2 , and so tearing occurs when the average depth of water on the sheet is $1.609/3 = .536 \text{ m}$ (about 21 inches).

If rain falls at the rate of .025 m (an inch) per hour, the sheet will tear in about 21 hours. Just before the sheet tears we have from (33) that

$$-\eta(1) = -\theta_0 = 1.260(.0447)^{1/3} = .447.$$

Since $\rho = 0.5$, Equation (32) gives

$$A = 1 - 1.399 (.0447)^{1/3} / .5 = .01.$$

The relative accuracy of this estimate for A is rather poor since the true value is $A \approx .09$. Despite this, the approximation is qualitatively correct in asserting that the sheet is almost completely engulfed when it tears. The depth of water is about 1.3 meters (4 feet) in the center of the puddle.

5. Conclusions

We conclude that this solution provides quite a bit of information about the behavior of wide flexible sheets during a rainstorm. The solution may also be useful in assessing the quality of finite-element codes that purport to solve geometrically nonlinear elastic problems. There is not at present a wide range of such solutions for comparison with finite-element results. The solution given here is not ideal for that purpose (some error is inherent in the numerical integration and solution by the Brent algorithm) but is probably good enough to provide a meaningful standard.

The example suggests that, if the assumed parameters are typical of those in common use, tearing will occur only in a phenomenally heavy or prolonged rainstorm. However, lesser storms may cause unacceptably large deflections at the center of the sheet.

With minor modifications this solution can be employed to study the more realistic case where the edges of the sheet are attached to the ends of elastic columns. The solution in that case is more complicated than here, but not as much so as might be expected, for the following reason. In the present problem the sheet has fixed ends, and the only inextensional solution is trivial. When the ends are allowed to move, a nontrivial inextensional solution exists and is not much more difficult to find than the present one.

References

1. Antman, S.S., "Multiple Equilibrium States of Nonlinearly Elastic Strings", SIAM Journal of Applied Mathematics, Volume 37, 1979, pp. 588-604.
2. Pugsley, A., "The Nonlinear Behavior of a Suspended Cable", Quarterly Journal of Mechanics and Applied Mathematics, Volume 30, pp. 157-162.
3. Fried, I., "Large Deformation Static and Dynamic Finite Element Analysis of Extensible Cables", Computers and Structures, Volume 15, 1982, pp. 315-319.
4. Huddleston, J.V. "Computer Analysis of Extensible Cables", Journal of the Engineering Mechanics Division, Proceeding of the American Society of Civil Engineers, Volume 107, 1981, pp. 27-37.
5. Peyrot, A.H. and Goulois, A.M. "Analysis of Cable Structures", Computers and Structures, Volume 10, 1979, pp. 805-813.
6. Malcolm, D.J. and Glockner, P.G. "Collapse by Ponding of Air-Supported Membranes", Journal of the Structural Division, Proceeding of the American Society of Civil Engineers, Volume 104, 1978, pp. 1525-1532.

Appendix

The objective is to find the asymptotic behavior as $A \rightarrow 1$ and $\theta_0 \rightarrow 0$ of the solutions to

$$A(Q + \cos\theta_0) + Q^{1/2} \rho^{-1} \pi/2 - 1 = 0 \quad (\text{A.1})$$

$$A + Q^{1/2} \rho^{-1} [(1+Q)^{-1} \pi/2 - I_x] - 1 = 0 \quad (\text{A.2})$$

where

$$I_x = \int_{\theta_0}^0 \psi(t, \theta_0) (Q + \cos t)^{-2} \sin t dt \quad (\text{A.3})$$

$$\psi(t, \theta_0) = (\pi/2) - \sin^{-1}(\tan t / \tan \theta_0) \quad (\text{A.4})$$

We start by estimating I_x . Since θ_0 is small and negative so is t , hence

$$Q + \cos t \approx 1 + Q - t^2/2 \approx 1 + Q \quad (\text{A.5})$$

$$\sin t \approx \tan t \approx t, \quad \sin \theta_0 \approx \theta_0$$

$$\begin{aligned} I_x &\approx (1+Q)^{-2} \int_{\theta_0}^0 [(\pi/2) - \sin^{-1}(t/\theta_0)] t dt \\ &\approx (1+Q)^{-2} [-(\pi\theta_0^2/4) - \int_{\theta_0}^0 t \sin^{-1}(t/\theta_0) dt] \end{aligned}$$

and

$$\int_{\theta_0}^0 t \sin^{-1}(t/\theta_0) dt = -\pi\theta_0^2/8.$$

Thus we find

$$I_x \approx - (1+Q)^{-2} \pi\theta_0^2/8. \quad (\text{A.6})$$

If the $t^2/2$ term were retained in equation (A.5), a negligible correction to this estimation would result.

Now we define

$$\gamma = 1 - A > 0, \quad Q^{1/2} = \alpha\beta > 0, \quad \alpha = 2\rho/\pi \quad (\text{A.7})$$

and use the approximation

$$\cos\theta_0 \approx 1 - \theta_0^2/2$$

in equation (A.1) to obtain

$$\theta_0^2/2 \approx [(\beta - \gamma)/(1 - \gamma)] + \alpha^2\beta^2$$

It is evident now that as $\theta_0 \rightarrow 0$ and $A \rightarrow 1$, i.e. $\gamma \rightarrow 0$, we must have $\beta \rightarrow 0$. We define

$$\phi = \beta - \gamma \quad (A.8)$$

where $|\phi| \ll \beta, \gamma$, and so can write

$$\theta_0^2/2 \approx [\phi/(1 - \gamma)] + \alpha^2\beta^2 \quad (A.9)$$

We do not yet know the relative magnitudes of ϕ and β^2 . To clarify this we observe that (A.2) can be written with the aid of (A.6) and (A.7) as

$$-\gamma(1 + \alpha\beta)^2 + \beta(1 + \alpha\beta) + \beta\theta_0^2/4 = 0.$$

Using equations (A.8) and (A.9) leads eventually to

$$2(\phi - 2\gamma\alpha^2\beta^2 + \alpha^2\beta^3) + \beta\phi/(1 - \gamma) = 0$$

which is solved for ϕ to obtain

$$\begin{aligned} \phi &\sim \alpha^2\gamma^3, \quad \beta \sim \gamma(1 + \alpha^2\gamma^2) \\ Q &\sim \alpha^2\gamma^2(1 + 2\alpha^2\gamma^2) \end{aligned} \quad (A.10)$$

Combining this with (A.9), the estimate

$$\theta_0^2/2 \sim \alpha^2\gamma^2 + \alpha^2\gamma^3/(1 - \gamma) \sim \alpha^2\gamma^2$$

$$\theta_0 \sim 2^{1/2}\alpha\gamma$$

is found.

The other quantities of interest are calculated now, as follows:

$$\epsilon_0 = Q/\cos\theta_0 \sim \alpha^2\gamma^2$$

$$\lambda = -\epsilon_0 \sin\theta_0 \sim 2^{1/2}\alpha^3\gamma^3$$

$$\delta = -A(1 + \epsilon_0) \sin\theta_0 \sim 2^{1/2}\alpha\gamma$$

$$\eta(1) = -\delta + Q^{1/2}\rho^{-1} \tan\theta_0 \sim -2^{1/2}\alpha\gamma$$

When these formulas are re-expressed in terms of λ , Equations (32), (33) and (34) are obtained.

Figure Captions

Figure 1: Sketch of the Deformed String

Figure 2: Load on an Element of the Deformed String

Figure 3: Deformed Shapes of the String for $\rho = 1/2$. Initial and Deformed Positions for $A = 0.4$ (o) and $A = 0.8$. (x).

Figure 4: Dependence of Strain, ϵ_0 , on Total Weight, λ , for $\rho = 0.2$, (o), 0.5 (x) and 0.8 (Δ).

Figure 5: Dependence of Angle, θ_0 , on Total Weight, λ , for $\rho = 0.2$, (o), 0.5 (x) and 0.8 (Δ).

Figure 6: Dependence of Central Deflection, $\eta(1)$, on Total Weight, λ , for $\rho = 0.2$, (o), 0.5 (x) and 0.8 (Δ).

Figure 7: Dependence of Puddle Edge Position, A , on Total Weight, λ , for $\rho = 0.2$ (o), 0.5 (x) and 0.8 (Δ).

Figure 8: Dependence of Puddle-Surface Displacement, δ , on Total Weight, λ , for $\rho = 0.2$ (o), 0.5 (x) and 0.8 (Δ).

Figure 9: Dependence of maximum Load λ_f on Parameter ρ .

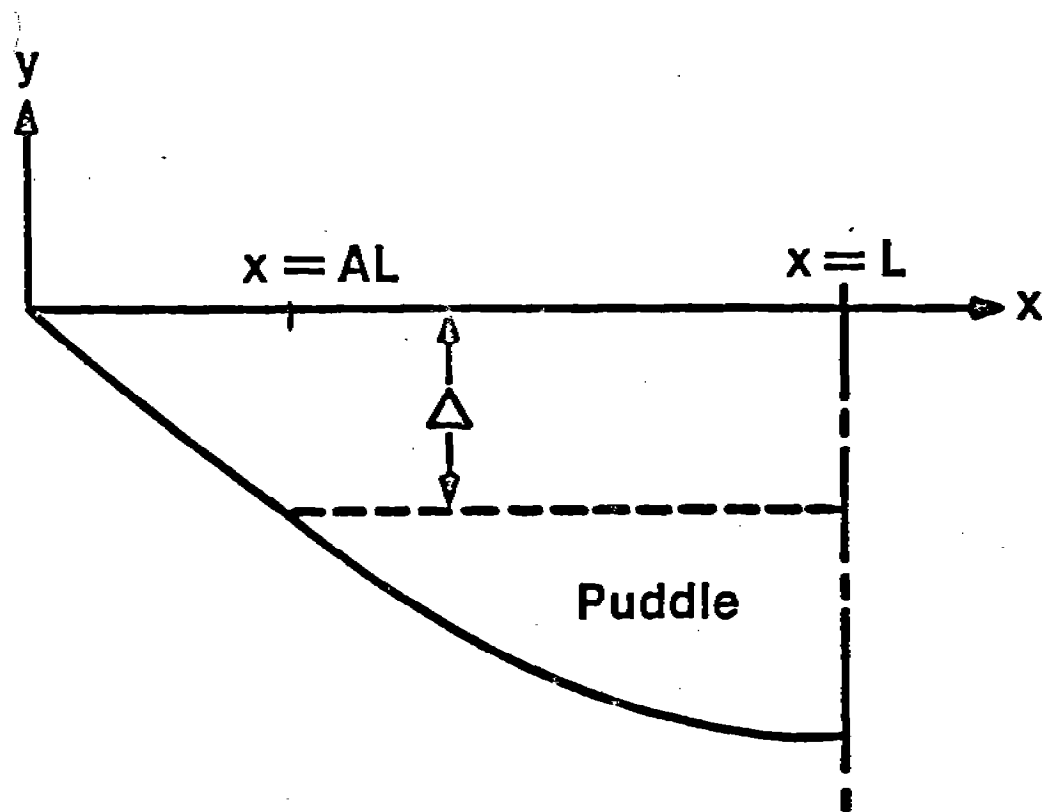


Fig 1

Figure 1: Sketch of the Deformed String

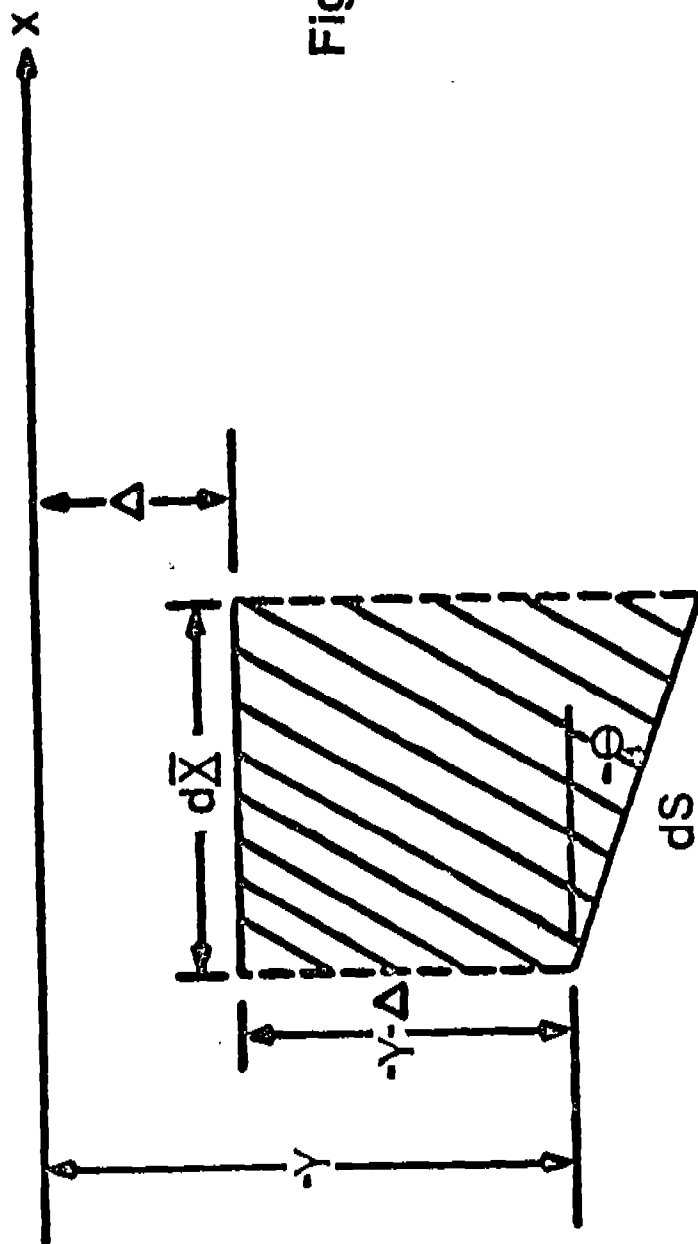


Fig 2

Figure 2: Load on an Element of the Deformed String

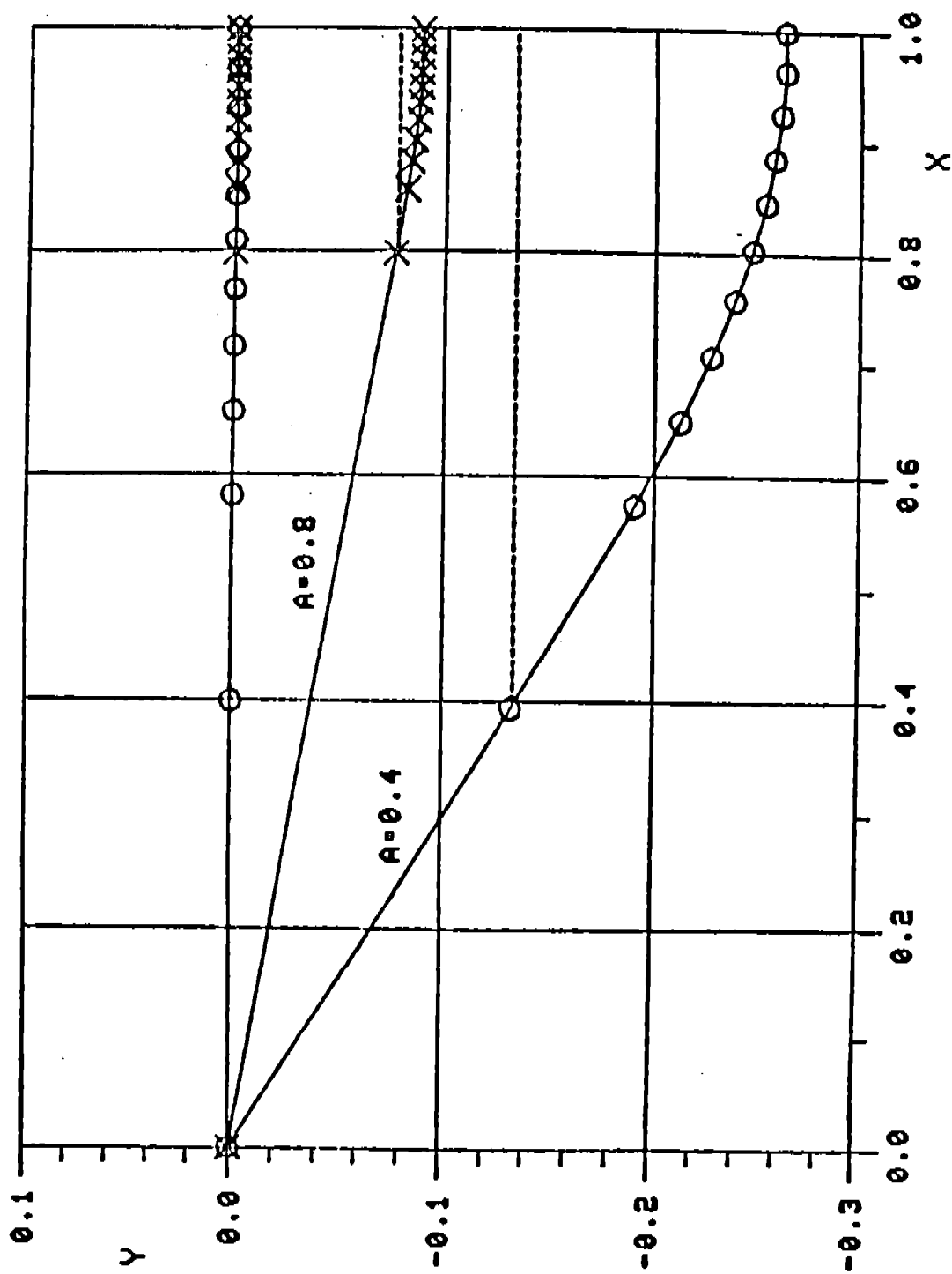


Figure 3: Deformed Shapes of the String for $\rho = 1/2$. Initial and Deformed Positions for $A = 0.4$ (o) and $A = 0.8$. (x).

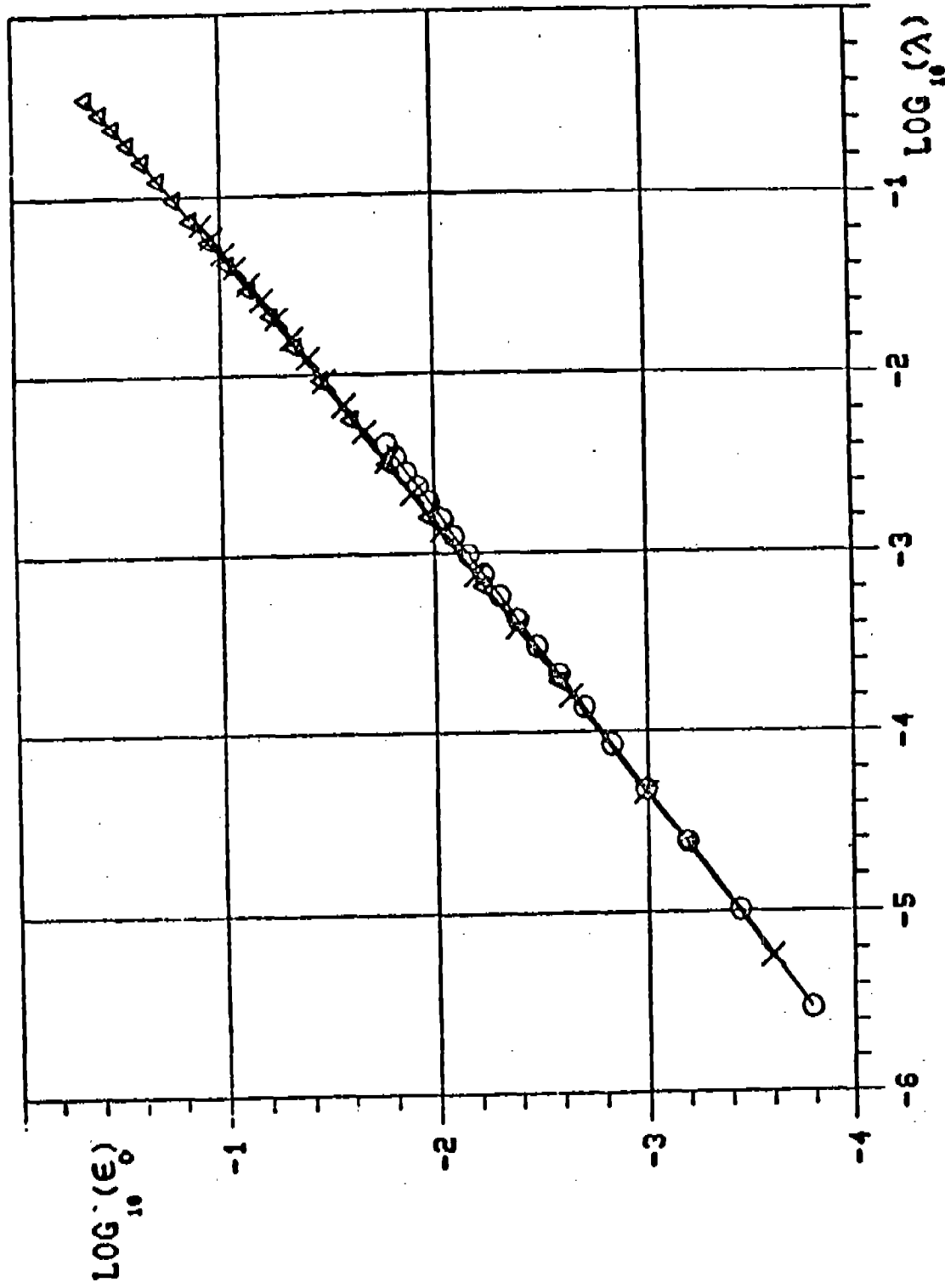


Figure 4: Dependence of Strain, ϵ_0 , on Total Weight, λ , for $p = 0.2$, \circ , 0.5 (\times) and 0.8 (Δ).

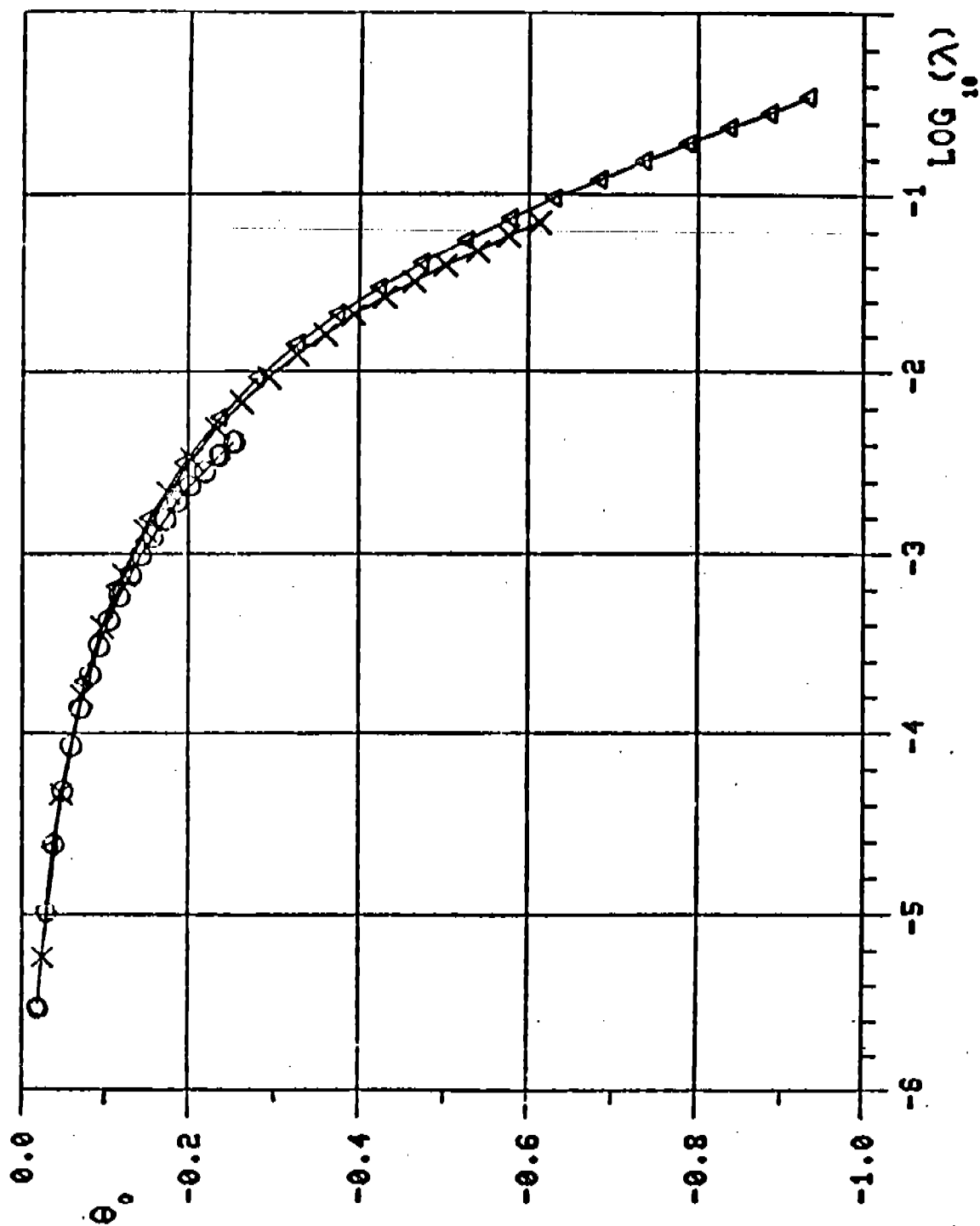


Figure 5: Dependence of Angle, θ_0 , on Total Weight, λ , for $\rho = 0.2$, (o), 0.5 (x) and 0.8 (Δ).

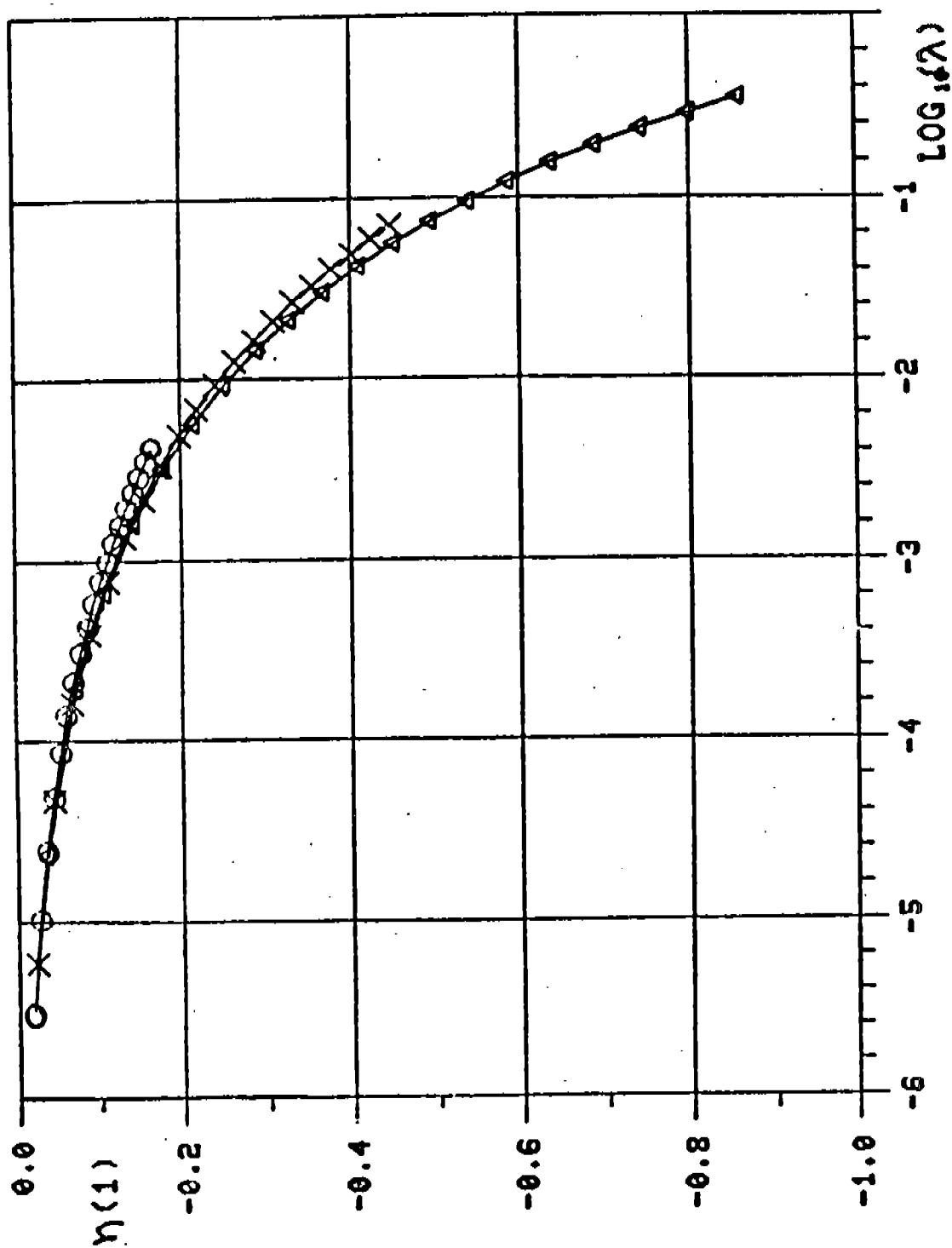


Figure 6: Dependence of Central Deflection, $\eta(1)$, on Total Weight, λ , for $\rho = 0.2$, (o), 0.5 (x) and 0.8 (Δ).

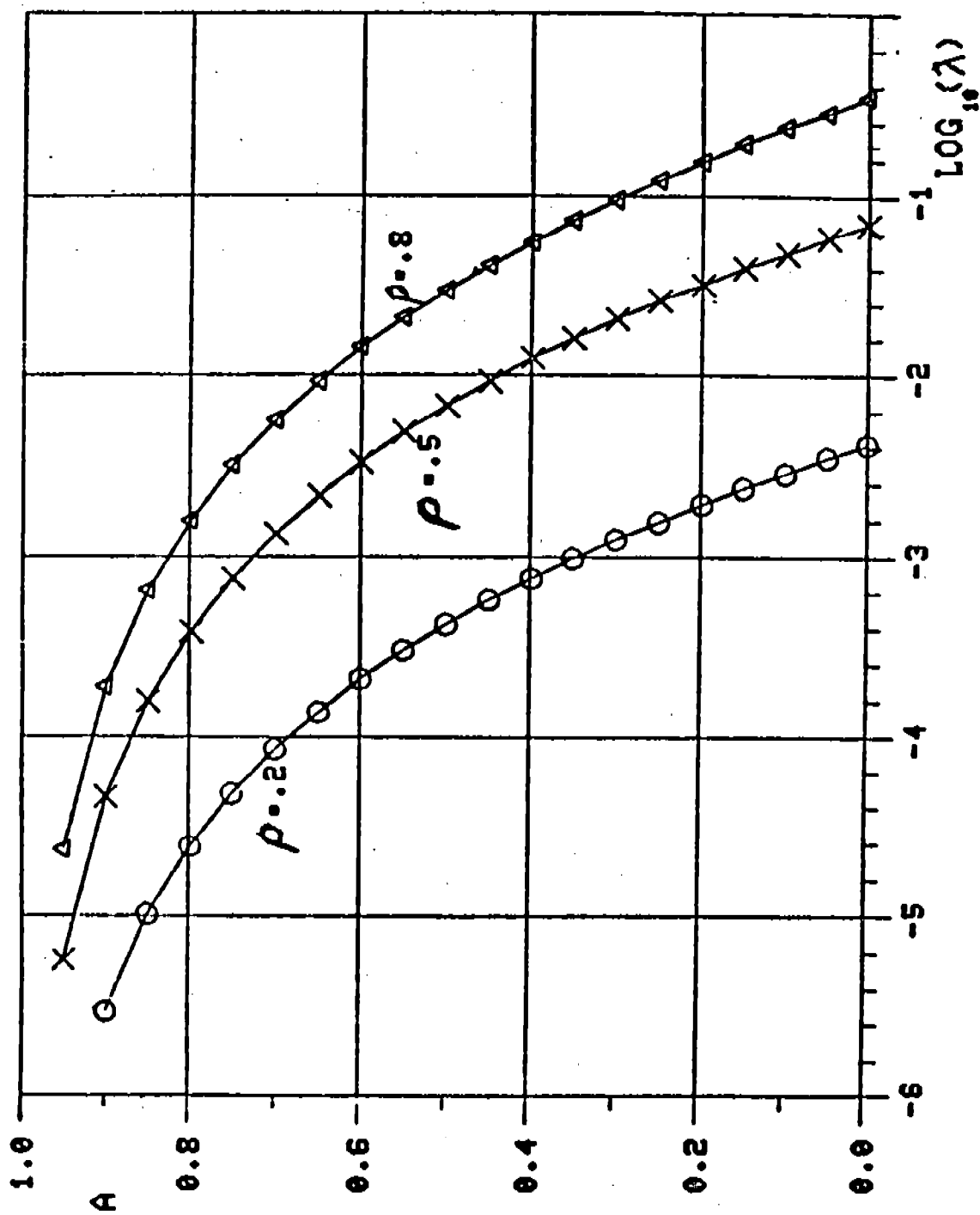


Figure 7: Dependence of Puddle Edge Position, A , on Total Weight, λ , for $p = 0.2$ (o), 0.5 (x) and 0.8 (Δ).

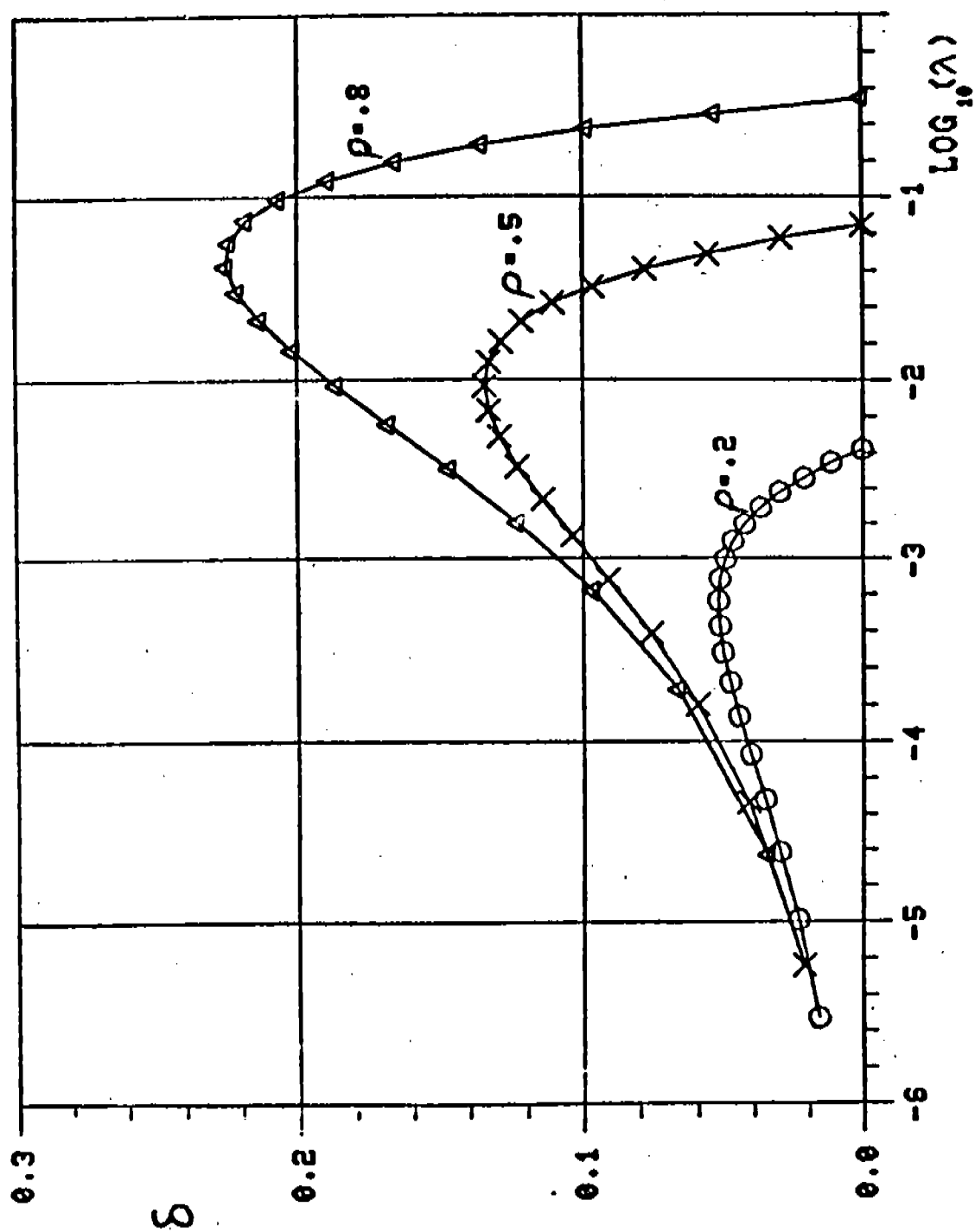


Figure 8: Dependence of Puddle-Surface Displacement, δ , on Total Weight, λ , for $\rho = 0.2$ (o), 0.5 (x) and 0.8 (Δ).

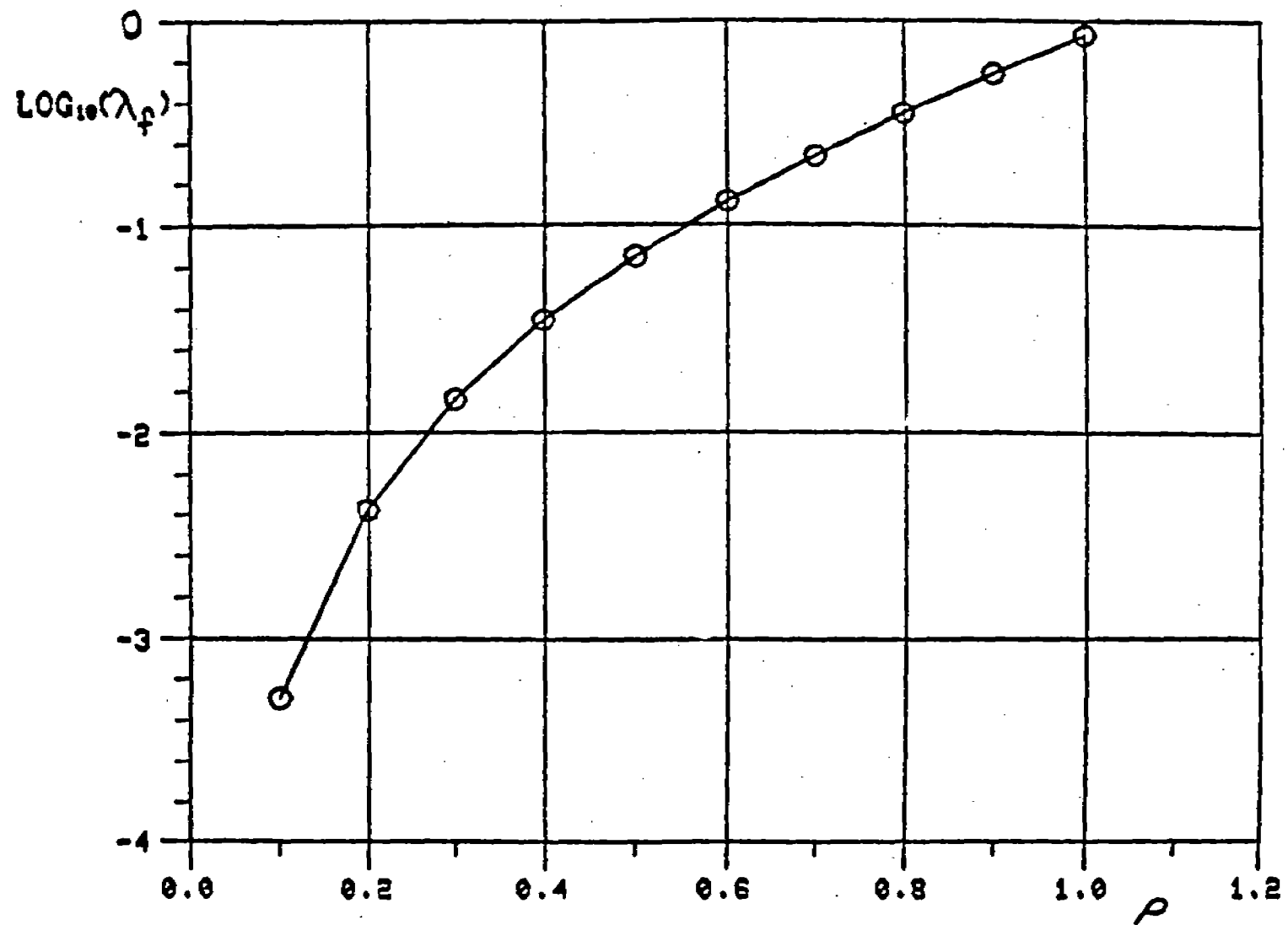


Figure 9: Dependence of maximum Load λ_f on Parameter ρ .

RELATIVISTIC WAVE EQUATIONS FOR SOLIDS
AND LOW TEMPERATURE QUANTUM SYSTEMS

Richard A. Weiss
Environmental Systems Division
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180-0631

ABSTRACT. The local scale invariance of relativistic thermodynamics is established and a set of coupled relativistic wave equations is developed that describe the propagation of small amplitude waves in solids and low temperature interacting Fermi and Bose systems. A first order expansion calculation is done in order to obtain the wave equations from a basic set of material equations. The wave equations determine the relativistic energy density and Grüneisen parameter for small amplitude mechanical or electromagnetic waves. In turn, the wave amplitude and phase velocity are obtained from the energy density and Grüneisen parameter of the waves. At low pressures the wave amplitudes are determined to be not greatly different from those predicted by a nonrelativistic calculation, but at high pressures, such as those expected to occur under nuclear blast loading or in stellar compact objects, the calculated relativistic wave amplitudes can be considerably larger than the corresponding nonrelativistic predictions.

1. INTRODUCTION. Local gauge invariance has become a powerful tool in modern physics.¹ It has unified the electromagnetic force with the weak nuclear force, and possibly also with the strong nuclear force, into a single entity described by local non-Abelian gauge invariance of the fields.² The requirement of local gauge invariance necessitates the introduction of new fields which obey sets of coupled differential equations which lead to unification of the fields in a natural way. The gauge transformations correspond to generalized rotations of the fields in the manner $e^{-i\phi(x)}$ where $\phi(x)$ depends on space and time for local gauge invariance.

A theory of relativistic thermodynamics has been developed for solids and quantum liquids which is based on a set of coupled differential equations for the zero temperature values of the Grüneisen parameter and pressure.³ These coupled equations are derived from a relativistic trace equation that has been gauged (scaled) by the introduction of the Grüneisen parameter. The fact that an additional field (the Grüneisen parameter) has to be considered in order to calculate the relativistic pressure suggests an invariance of the system similar to local gauge invariance. Electromagnetism and the electro-weak interactions are examples of fields that are described by local gauge theories in which the gauging is accomplished by the introduction of additional fields. The invariance of relativistic thermodynamics refers to different values of pressure and energy density; and, in analogy to a local gauge transformation, is manifested through a real exponential $e^{-\phi(V,T)}$ corresponding to changes of the local scale of the pressure and energy density of a system.⁴ This paper shows that relativistic thermodynamics is scale invariant, and that the local symmetry group for relativistic thermodynamics

is the unimodular group of scale transformations $e^{-\phi(V,T)}$.

The trace equation of relativistic thermodynamics is written as³

$$U + T \left(\frac{dU}{dT} \right)_{PV} - 3V \frac{d}{dV}(PV)_U = U^a + T \left(\frac{dU^a}{dT} \right)_{P^a V} \quad (1)$$

where U = relativistic internal energy, P = relativistic pressure, T = absolute temperature, V = volume per mole of substance, and U^a and P^a = corresponding nonrelativistic internal energy and pressure. Throughout this paper the index "a" will refer to nonrelativistic calculations. It has been shown that for a physical system described by $U = U_0 + U_j T^j + \dots$ and $P = P_0 + P_j T^j + \dots$ (such as the high temperature Mie-Grüneisen state equation with $j = 1$ and the Debye state equation with $j = 4$) the trace equation (1) is equivalent to the following set of coupled differential equations³

$$3V^2 \frac{d^2 P_0}{dV^2} + 3(3 + \gamma_0)V \frac{dP_0}{dV} + [3(\gamma_0 + V \frac{d\gamma_0}{dV}) + 4]P_0 = P_0^a \quad (2)$$

$$U_j \left(1 + j + \frac{j\gamma_0 P_0}{P_0 - K_0} - 3V \frac{d\gamma_0}{dV} \right) = U_j^a \left(1 + j + \frac{j\gamma_0^a P_0^a}{P_0^a - K_0^a} \right) \quad (3)$$

where the internal energy coefficients are given by⁵

$$\frac{U_j^a}{U_j} = \exp \left[(j - 1) \int^V (\gamma_0^a - \gamma_0) \frac{dV}{V} \right] \quad (4)$$

and where P_0 , K_0 and γ_0 = zero temperature values of the relativistic pressure, incompressibility ($-VdP_0/dV$) and Grüneisen parameter respectively, and P_0^a , K_0^a and γ_0^a are the corresponding nonrelativistic values of these quantities. Eqs. (2) and (3) are a set of coupled nonlinear differential equations for P_0 and γ_0 . The zero temperature value of the Grüneisen parameter is the $T = 0$ limit of the Grüneisen parameter defined as⁵

$$\gamma = \frac{V}{C_V} \left(\frac{\partial P}{\partial T} \right)_V \quad (5)$$

which for a temperature dependent Debye temperature becomes,^{6,7}

$$\gamma = \frac{-\frac{V}{\theta_D} \left(\frac{\partial \theta_D}{\partial V} \right)_T}{\left[1 - \frac{T}{\theta_D} \left(\frac{\partial \theta_D}{\partial T} \right)_V \right]} \quad (6)$$

where C_V = heat capacity and θ_D = Debye temperature. The Debye temperature is generally a function of V and T . Equation (6) gives the general nonrelativistic expression for the Grüneisen parameter.

Wave propagation in matter is complicated by the fact that matter is a thermodynamic system which is described by physical quantities such as the heat capacity and Grüneisen parameter.^{5,8} Further complications arise because matter is often prestressed as for instance by gravitation in a star or planet.⁹ Various assumptions are used to calculate the phase velocity and wave amplitude in terms of the thermodynamic state equation of a material medium, but no completely general procedure exists for calculating these quantities. This paper develops a relativistic calculation of the phase velocity and amplitude for mechanical waves in thermodynamic media.

Specifically, this paper develops a set of coupled first order relativistic equations that govern small amplitude wave propagation in thermodynamic systems such as solids and low temperature Fermi and Bose liquids. These equations determine the relativistic energy density and Grüneisen parameter for mechanical or electromagnetic waves. The phase velocity and wave amplitude can be obtained from the energy density and Grüneisen parameter for the waves, and are expressed in terms of the material parameters of the thermodynamic ground state of the material system.

The wave equations are developed from a small amplitude perturbation expansion of a set of relativistic material equations. This insures that the derived wave equations are intimately connected to the material parameters of the thermodynamic medium. A complicated set of nonlinear wave equations are derived whose complete solution requires numerical computer techniques, however, an order of magnitude approximate solution is given for the case of elastic waves in solids. The procedure followed in this paper is to first review local scale invariance, then the relativistic ground state calculation, and then the relativistic wave equations that govern excitations in a thermal medium.

2. LOCAL SCALE INVARIANCE. The fact that an additional field $\gamma_0(V)$ must be introduced in order to calculate the zero temperature pressure $P_0(V)$ as in Eqs. (2) and (3) suggests that the relativistic trace equation (1) may be invariant under a local scale transformation of the form $\exp[-\phi(V,T)]$. The scale invariance of Eq. (1) will now be demonstrated. To do this the following elementary thermodynamic relationships are used³

$$\left(\frac{dU}{dT}\right)_{PV} = \left(\frac{\partial U}{\partial T}\right)_V - \frac{V}{(P - K_T)} \left(\frac{\partial U}{\partial V}\right)_T \left(\frac{\partial P}{\partial T}\right)_V \quad (7)$$

$$\frac{d}{dV}(PV)_U = P - K_T - \gamma \left[T \left(\frac{\partial P}{\partial T}\right)_V - P\right] \quad (8)$$

where $K_T = -V(\partial P/\partial V)_T$ = isothermal incompressibility. Using Eqs. (7) and (8) allows Eq. (1) to be written as

$$\begin{aligned} & \left(1 + T \frac{\partial}{\partial T} - bV \frac{\partial}{\partial V}\right) U - 3V \left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T}\right) P \\ & = \left(1 + T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V}\right) U^a \end{aligned} \quad (9)$$

where

$$b = \frac{T(\partial P/\partial T)_V}{(P - K_T)} \quad (10)$$

$$b^a = \frac{T(\partial P^a/\partial T)_V}{(P^a - K_T^a)} \quad (11)$$

Note that the Grüneisen parameter γ defined in equation (5) is not independent of the quantity b defined by equation (10). Eq. (9) can be rewritten in a more symmetrical form by writing $U = EV$, where E = energy density, with the result

$$\begin{aligned} & \left(1 - b + T \frac{\partial}{\partial T} - bV \frac{\partial}{\partial V}\right) E - 3 \left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T}\right) P \\ & = \left(1 - b^a + T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V}\right) E^a \end{aligned} \quad (12)$$

Eq. (12) can be written as

$$(H_1 - W_1)E - 3(H_2 - W_2)P = (H_1^a - W_1^a)E^a \quad (13)$$

where

$$H_1 = T \frac{\partial}{\partial T} - bV \frac{\partial}{\partial V} \quad (14)$$

$$H_2 = V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T} \quad (15)$$

$$W_1 = b - 1 \quad (16)$$

$$W_2 = -\gamma - 1 \quad (17)$$

Eq. (13) is in a form suitable for demonstrating local scale invariance.

Introduce the following scale transformations $E \rightarrow E' = Ee^{-\psi}$ and $P \rightarrow P' = Pe^{-\phi}$ where ψ and ϕ are functions of V and T . Also for the nonrelativistic energy density introduce the scale transformation $E^a \rightarrow E^{a'} = E^a e^{-\psi^a}$ where ψ^a is also a function of V and T . Under these scale transformations γ and b assume new values $\gamma \rightarrow \gamma'$ and $b \rightarrow b'$, and equations (14) through (17) become

$$H'_1 = T \frac{\partial}{\partial T} - b' V \frac{\partial}{\partial V} \quad (18)$$

$$H'_2 = V \frac{\partial}{\partial V} - \gamma' T \frac{\partial}{\partial T} \quad (19)$$

$$W'_1 = b' - 1 \quad (20)$$

$$W'_2 = -\gamma' - 1 \quad (21)$$

where b' and γ' are to be determined by the local scale invariance conditions. The local scale invariant conditions for the operators in equation (13) are written in a manner similar to local gauge invariance as²

$$(H'_1 - W'_1)Ee^{-\psi} = e^{-\psi}(H_1 - W_1)E \quad (22)$$

$$(H'_2 - W'_2)Pe^{-\phi} = e^{-\phi}(H_2 - W_2)P \quad (23)$$

$$(H_1^{a'} - W_1^{a'})E^a e^{-\psi^a} = e^{-\psi^a}(H_1^a - W_1^a)E^a \quad (24)$$

Equations (22) through (24) allows the trace equation (13) to be written as

$$e^{\psi}(H'_1 - W'_1)e^{-\psi}E - 3e^{\phi}(H'_2 - W'_2)e^{-\phi}P = e^{\psi^a}(H_1^{a'} - W_1^{a'})e^{-\psi^a}E^a \quad (25)$$

so that the following operator equations hold

$$H_1 - W_1 = e^\psi (H'_1 - W'_1) e^{-\psi} \quad (26)$$

$$H_2 - W_2 = e^\phi (H'_2 - W'_2) e^{-\phi} \quad (27)$$

The values of b' and γ' are obtained as follows. Placing equations (14) through (21) into the scale invariant conditions given by equations (22) and (23) yields the following results

$$b' = b + \frac{E \left(bV \frac{\partial \psi}{\partial V} - T \frac{\partial \psi}{\partial T} \right)}{E + V \frac{\partial E}{\partial V} - EV \frac{\partial \psi}{\partial V}} \quad (28)$$

$$\gamma' = \gamma + \frac{P \left(V \frac{\partial \phi}{\partial V} - \gamma T \frac{\partial \phi}{\partial T} \right)}{P - T \frac{\partial P}{\partial T} + PT \frac{\partial \phi}{\partial T}} \quad (29)$$

Now since b' and γ' are associated E' and P' respectively, while b and γ are associated with E and P respectively, one can write

$$b' = b + \frac{db}{dE} (E' - E) + \dots \quad (30)$$

$$\gamma' = \gamma + \frac{d\gamma}{dP} (P' - P) + \dots \quad (31)$$

while from $E' = Ee^{-\psi}$ and $P' = Pe^{-\phi}$ it follows that

$$E' - E = E(e^{-\psi} - 1) \quad (32)$$

$$P' - P = P(e^{-\phi} - 1) \quad (33)$$

so that combining equations (28) through (33) yields the following differential equations for ψ and ϕ ,

$$\frac{db}{dE} = \frac{e^\psi}{e^\psi - 1} \left(\frac{T \frac{\partial \psi}{\partial T} - bV \frac{\partial \psi}{\partial V}}{E + V \frac{\partial E}{\partial V} - EV \frac{\partial \psi}{\partial V}} \right) \quad (34)$$

$$\frac{d\gamma}{dP} = \frac{e^\phi}{e^\phi - 1} \left(\frac{\gamma T \frac{\partial \phi}{\partial T} - V \frac{\partial \phi}{\partial V}}{P - T \frac{\partial P}{\partial T} + PT \frac{\partial \phi}{\partial T}} \right) \quad (35)$$

Therefore it is always possible to find a ψ and ϕ such that the trace equation (13) is locally scale invariant.

Consider now the infinitesimal local scale transformation where ψ and ϕ are small quantities, then equations (32) and (33) become

$$E' - E = -\psi E \quad (36)$$

$$P' - P = -\phi P \quad (37)$$

and equations (34) and (35) become

$$\frac{db}{dE} = \frac{\frac{T}{\psi} \frac{\partial \psi}{\partial T} - b \frac{V}{\psi} \frac{\partial \psi}{\partial V}}{E + V \frac{\partial E}{\partial V} - EV \frac{\partial \psi}{\partial V}} \quad (38)$$

$$\frac{d\gamma}{dP} = \frac{\gamma \frac{T}{\phi} \frac{\partial \phi}{\partial T} - \frac{V}{\phi} \frac{\partial \phi}{\partial V}}{P - T \frac{\partial P}{\partial T} + PT \frac{\partial \phi}{\partial T}} \quad (39)$$

Therefore it is always possible to find a ψ and ϕ such that the trace equation (13) is invariant under an infinitesimal local scale transformation. Note that for the case when b and γ are slowly varying functions, equations (38) and (39) yield

$$b = \frac{T \frac{\partial \psi}{\partial T}}{V \frac{\partial \psi}{\partial V}} \quad (40)$$

$$\gamma = \frac{V \frac{\partial \phi}{\partial V}}{T \frac{\partial \phi}{\partial T}} \quad (41)$$

It turns out that the solutions to equations (40) and (41) are very simple. Choose ψ and ϕ as follows

$$\psi = \frac{PV}{P_1 V_1} \quad (42)$$

$$\phi = \frac{\theta_D}{T} \quad (43)$$

where $P_1 V_1$ = an initial value of PV . Then from equations (40) and (41) it follows that

$$b = \frac{T \frac{\partial P}{\partial T}}{(P - K_T)} \quad (44)$$

$$\gamma = \frac{-\frac{V}{\theta_D} \frac{\partial \theta_D}{\partial V}}{1 - \frac{T}{\theta_D} \frac{\partial \theta_D}{\partial T}} \quad (45)$$

Equation (44) is just the basic definition of the quantity b given in equation (10), while equation (45) is just the standard result for the Grüneisen parameter given in equation (6). Therefore the values of ψ and ϕ given by equations (42) and (43) are the proper potential functions for the infinitesimal local scale transformation of relativistic thermodynamics for the case of slowly varying values of b and γ . The explicit determination of ψ and ϕ for the general case of local scale invariance given by equations (34) and (35) is much more complicated, and has not been accomplished.

The scale invariance of the trace equations (1) or (9) requires the presence of the parameters b and γ , although as mentioned earlier, b and γ are not independent. This means that for the zero-temperature case, the zero-temperature values of the pressure and the Grüneisen parameter must be determined simultaneously as shown in equations (2) and (3). The establishment of local scale invariance of the relativistic trace equation gives one confidence to use equations (2) and (3) to determine the equations governing the propagation of small amplitude waves in a thermodynamic medium. But first the ground state of the thermodynamic system must be described.

3. GROUND STATE. The nonrelativistic state equation of the ground state of a thermal system is assumed to have the following form⁵

$$E^a = E_o^a + E_j^a T^j + \dots \quad (46)$$

$$P^a = P_o^a + P_j^a T^j + \dots \quad (47)$$

where E^a and P^a = nonrelativistic energy density and pressure respectively, E_0^a and P_0^a = nonrelativistic zero-temperature values of the energy density and pressure respectively, E_j^a and P_j^a = nonrelativistic thermal coefficients for the energy density and pressure respectively, T = absolute temperature of the system ($^{\circ}\text{K}$), and j = numerical index having values characteristic of the type of physical system. Typical examples of systems that are described by equations (46) and (47) are⁵

- $j = 1$ high temperature solid
- $j = 2$ low temperature Fermi gas
- $j = 5/2$ low temperature molecular Bose gas
- $j = 4$ low temperature solid

A commonly used descriptor of the thermal state equations given by equations (46) and (47) is the nonrelativistic zero-temperature value of the Grüneisen parameter that is defined by⁵

$$\gamma_0^a = \frac{P_j^a}{E_j^a} = \frac{1}{(j-1)} \frac{1}{E_j^a} \frac{d}{dV} (VE_j^a) \quad (48)$$

except for $j = 1$. Here γ_0^a = nonrelativistic zero-temperature value of the Grüneisen parameter, and V = volume of the material system. When $j = 1$, $\gamma_0^a = 2/3$.

The corresponding relativistic state equations will be written as

$$E = E_0 + E_j T^j + \dots \quad (49)$$

$$P = P_0 + P_j T^j + \dots \quad (50)$$

$$\gamma_0 = \frac{P_j}{E_j} = \frac{1}{(j-1)} \frac{1}{E_j} \frac{d}{dV} (VE_j) \quad (51)$$

except for $j = 1$, when $\gamma_0 = 2/3$, and where E_0 and P_0 = relativistic zero-temperature energy density and pressure respectively, E_j and P_j = relativistic thermal coefficients for the energy density and pressure respectively, and γ_0 = relativistic zero-temperature Grüneisen parameter.

The relativistic values of the zero-temperature energy density and Grüneisen parameter for the ground state thermodynamic system are given by the solution of the following two coupled equations³

$$E_0 - 3[(1 + \gamma_0)P_0 - K_0] = E_0^a \quad (52)$$

$$E_j \left(1 + j + \frac{j\gamma_o P_o}{P_o - K_o} + 3n \frac{d\gamma_o}{dn} \right) = E_j^a \left(1 + j + \frac{j\gamma_o^a P_o^a}{P_o^a - K_o^a} \right) \quad (53)$$

where $n = 1/V$ = reciprocal volume

$K_o = n \frac{dP_o}{dn}$ = relativistic zero-temperature bulk modulus

$K_o^a = n \frac{dP_o^a}{dn}$ = nonrelativistic zero-temperature bulk modulus

From equations (48) through (51) it follows that³

$$E_j = nC_j \exp \left[- (j-1) \int^n \gamma_o \frac{dn}{n} \right] \quad (54)$$

$$E_j^a = nC_j^a \exp \left[- (j-1) \int^n \gamma_o^a \frac{dn}{n} \right] \quad (55)$$

so that

$$\frac{E_j}{E_j^a} = \frac{C_j}{C_j^a} \exp \left[- (j-1) \int^n (\gamma_o - \gamma_o^a) \frac{dn}{n} \right] \quad (56)$$

When $\gamma_o = \gamma_o^a$, it must follow that $E_j = E_j^a$ so that quite generally $C_j = C_j^a$. Combining equation (56) with equation (53) shows that equations (52) and (53) are two nonlinear coupled equations for determining E_o and γ_o in terms of the known values of E_o^a and γ_o^a . These equations give the recipe for calculating the relativistic ground state of a thermal system in terms of the corresponding nonrelativistic description of the ground state. The calculation of the relativistic excited states will now be given.

4. EXCITATIONS. The excitations in thermal media that are considered in this paper are mechanical radiation and electromagnetic waves. Only waves of small amplitude are treated. The thermal state equations of the radiation are assumed to have a form similar to the ground state equations (46), (47), (49) and (50) and are written as

$$E_r^a = E_{or}^a + E_{jr}^a T^j + \dots \quad (57)$$

$$P_r^a = P_{or}^a + P_{jr}^a T^j + \dots \quad (58)$$

and

$$E_r = E_{or} + E_{jr} T^j + \dots \quad (59)$$

$$P_r = P_{or} + P_{jr} T^j + \dots \quad (60)$$

where

E_{or}^a and P_{or}^a = nonrelativistic zero-temperature radiation energy density and pressure respectively

E_{jr}^a and P_{jr}^a = nonrelativistic thermal coefficients for the radiation energy density and pressure respectively

E_{or} and P_{or} = relativistic zero-temperature radiation energy density and pressure respectively

E_{jr} and P_{jr} = relativistic thermal coefficients for the radiation energy density and pressure respectively

It will be shown in this paper that the relativistic state equations for radiation given in equations (59) and (60) must have nonzero thermal components even when the corresponding nonrelativistic thermal components that appear in equations (57) and (58) are taken to be zero. Finally, the radiation terms are assumed to be much smaller than the ground state terms, i.e., $E_r \ll E$ and $P_r \ll P$.

When radiation is present in a system whose ground state is described by P_o , K_o , and γ_o , these three parameters become $P_o + P_{or}$, $K_o + K_{or}$, and $\gamma_o + \delta_{or}$, where K_{or} = relativistic zero-temperature incompressibility associated with the radiation, and δ_{or} = incremental change in the relativistic Grüneisen parameter of the system due to the presence of radiation. The bulk modulus (incompressibility) associated with the radiation is given by

$$K_{or} = n \frac{dP_{or}}{dn} \quad (61)$$

The increment in the zero-temperature Grüneisen parameter of the system due to the presence of small amplitude radiation is obtained from the defining equation (51) by noting that when radiation is present this equation becomes

$$\gamma_o + \delta_{or} = \frac{P_j + P_{jr}}{E_j + E_{jr}} = \frac{P_j + P_{jr}}{E_j \left(1 + \frac{E_{jr}}{E_j} \right)} \quad (62)$$

Expanding the denominator in equation (62), keeping only first order terms, and finally subtracting equation (51) gives

$$\delta_{or} = \frac{E_{jr}}{E_j} \left(\frac{P_{jr}}{E_{jr}} - \frac{P_j}{E_j} \right) \quad (63)$$

$$= \frac{E_{jr}}{E_j} (\gamma_{or} - \gamma_o)$$

where γ_{or} = zero-temperature Grüneisen parameter for the radiation field itself, and is defined by

$$\gamma_{or} = \frac{P_{jr}}{E_{jr}} \quad (64)$$

The corresponding expressions for the nonrelativistic bulk modulus and Grüneisen parameters associated with the radiation field in matter are

$$K_{or}^a = n \frac{dP_{or}^a}{dn} \quad (65)$$

$$\delta_{or}^a = \frac{E_{jr}^a}{E_j^a} (\gamma_{or}^a - \gamma_o^a) \quad (66)$$

$$\gamma_{or}^a = \frac{P_{jr}^a}{E_{jr}^a} \quad (67)$$

From equations (63) and (66) it follows that if E_{jr} and E_{jr}^a are small quantities then so also are δ_{or} and δ_{or}^a . However, the radiation Grüneisen parameters themselves, γ_{or} and γ_{or}^a , are generally not small quantities being the ratio of two small quantities, and at low pressures have the value 1/3 for isotropic radiation.⁵ The quantities E_{or} , P_{or} , K_{or} , E_{jr} , P_{jr} , and δ_{or} , and their nonrelativistic analogs, are taken to be small quantities. Finally, from equation (64) and the law of energy conservation (see Appendix A), it follows that the ratio of thermal terms that occurs in equation (63) can be written as

$$\frac{E_{jr}}{E_j} = \frac{D_{jr}}{C_j} \exp \left[- (j-1) \int^n (\gamma_{or} - \gamma_o) \frac{dn}{n} \right] \quad (68)$$

where D_{jr} = constant associated with radiation field.

5. DERIVATION OF WAVE EQUATIONS. When radiation is present in the thermodynamic systems being considered, equation (52) becomes

$$E_o + E_{or} - 3[(1 + \gamma_o + \delta_{or})(P_o + P_{or}) - (K_o + K_{or})] = E_o^a + E_{or}^a \quad (69)$$

Similarly, when radiation is present, equation (53) becomes

$$\begin{aligned} (E_j + E_{jr}) \left[1 + j + \frac{j(\gamma_o + \delta_{or})(P_o + P_{or})}{P_o + P_{or} - K_o - K_{or}} + 3n \frac{d}{dn} (\gamma_o + \delta_{or}) \right] \\ = (E_j^a + E_{jr}^a) \left[1 + j + \frac{j(\gamma_o^a + \delta_{or}^a)(P_o^a + P_{or}^a)}{P_o^a + P_{or}^a - K_o^a - K_{or}^a} \right] \end{aligned} \quad (70)$$

Equations (69) and (70) are the relativistic equations for waves in matter, however they are too complex for simple solutions to be obtained. Simplification can be achieved by assuming the radiation components to be small quantities.

Considering only first order terms in equation (69), and subtracting equation (52), gives

$$E_{or} - 3[(1 + \gamma_o)P_{or} - K_{or}] - 3P_o \delta_{or} = E_{or}^a \quad (71)$$

In a similar fashion, by expanding the denominators in equation (70), using $1/(1+x) \sim 1-x$, and keeping only first order terms, and finally subtracting equation (53) yields the following result

$$jE_j(\alpha K_{or} - \beta P_{or}) + 3E_j n \frac{d\delta_{or}}{dn} + \frac{jE_j P_o \delta_{or}}{P_o - K_o} \quad (72)$$

$$+ E_{jr} \left(1 + j + \frac{j\gamma_o P_o}{P_o - K_o} + 3n \frac{d\gamma_o}{dn} \right)$$

$$= jE_j^a (\alpha^a K_{or}^a - \beta^a P_{or}^a) + \frac{jE_j^a P_o^a \delta_{or}^a}{P_o^a - K_o^a} + E_{jr}^a \left(1 + j + \frac{j\gamma_o^a P_o^a}{P_o^a - K_o^a} \right)$$

where

$$\alpha = \frac{\gamma_o P_o}{(P_o - K_o)^2} \quad \alpha^a = \frac{\gamma_o^a P_o^a}{(P_o^a - K_o^a)^2} \quad (73)$$

$$\beta = \frac{\gamma_o K_o}{(P_o - K_o)^2} \quad \beta^a = \frac{\gamma_o^a K_o^a}{(P_o^a - K_o^a)^2} \quad (74)$$

The radiation equations (71) and (72) will now be written in a much simpler form.

Equation (63) can be used for δ_{or} that occurs in equations (71) and (72), while the first derivative of δ_{or} that occurs in equation (72) is evaluated by using equations (63) and (68) and can be written as (see Appendix B)

$$n \frac{d\delta_{or}}{dn} = \frac{E_{jr}}{E_j} \left[n \frac{d\gamma_{or}}{dn} - n \frac{d\gamma_o}{dn} - (j-1)(\gamma_{or} - \gamma_o)^2 \right] \quad (75)$$

Substituting the results of equations (63) and (75) into the two basic equations (71) and (72) yields

$$E_{or} - 3[(1 + \gamma_o)P_{or} - K_{or}] - 3 \frac{E_{jr}}{E_j} P_o (\gamma_{or} - \gamma_o) = E_{or}^a \quad (76)$$

and

$$jE_j(\alpha K_{or} - \beta P_{or}) + E_{jr} \left[1 + j + \frac{jP_o \gamma_{or}}{P_o - K_o} + 3n \frac{d\gamma_{or}}{dn} - 3(j-1)(\gamma_{or} - \gamma_o)^2 \right] \quad (77)$$

$$= jE_j^a \left(\alpha^a K_{or}^a - \beta^a P_{or}^a \right) + E_{jr}^a \left(1 + j + \frac{jP_o^a \gamma_{or}^a}{P_o^a - K_o^a} \right)$$

These two radiation equations can be further simplified by using some basic mathematical properties of the radiation field.

For radiation, the zero-temperature pressure is related to the zero-temperature energy density through the radiation Grüneisen parameter as follows (see Appendix A)

$$P_{or} = \gamma_{or} E_{or} \quad (78)$$

This proportionality of the pressure and energy density is characteristic of radiation fields. The bulk modulus of the radiation field can be written using equation (78) as

$$K_{or} = n \frac{dP_{or}}{dn} = \gamma_{or} n \frac{dE_{or}}{dn} + E_{or} n \frac{d\gamma_{or}}{dn} \quad (79)$$

Equations analogous to (78) and (79) hold for the nonrelativistic quantities P_{or}^a and K_{or}^a respectively. Placing equations (78) and (79) into equation (76) and (77) and then dividing equation (77) by E_j yields the two fundamental first order coupled nonlinear differential equations for E_{or} and γ_{or} ,

$$In \frac{dE_{or}}{dn} + JE_{or} + g = E_{or}^a \quad (80)$$

$$Gn \frac{dE_{or}}{dn} + RE_{or} + f = \psi_{or}^a \quad (81)$$

$$\text{where } I = 3\gamma_{or} \quad (82)$$

$$J = 3n \frac{d\gamma_{or}}{dn} - 3(1 + \gamma_o)\gamma_{or} + 1 \quad (83)$$

$$g = 3 \frac{E_{jr}}{E_j} P_o (\gamma_o - \gamma_{or}) \quad (84)$$

$$G = j\alpha\gamma_{or} \quad (85)$$

$$R = j \left(\alpha n \frac{d\gamma_{or}}{dn} - \beta\gamma_{or} \right) \quad (86)$$

$$f = \frac{E_{jr}}{E_j} \left[1 + j + \frac{jP_o\gamma_{or}}{P_o - K_o} + n \frac{d\gamma_{or}}{dn} - 3(j-1)(\gamma_{or} - \gamma_o)^2 \right] \quad (87)$$

$$\psi_{or}^a = \frac{E_{jr}^a}{E_j} \left[1 + j + \frac{jP_o^a\gamma_{or}^a}{P_o^a - K_o^a} \right] + G^a n \frac{dE_{or}^a}{dn} + R^a E_{or}^a \quad (88)$$

$$G^a = j \frac{E_j^a}{E_j} \alpha^a \gamma_{or}^a \quad (89)$$

$$R^a = j \frac{E_j^a}{E_j} \left(\alpha^a n \frac{d\gamma_{or}^a}{dn} - \beta^a \gamma_{or}^a \right) \quad (90)$$

The ratio E_{jr}/E_j that occurs in equation (84) and (87) is a function of γ_{or} as shown in equation (68), while the ratio E_{jr}^a/E_j that appears in equations (89) and (90) is given by equation (56). Equations (80) and (81) are a set of coupled relativistic wave equations that determine the relativistic energy density E_{or} and Grüneisen parameter γ_{or} for radiation in terms of the corresponding nonrelativistic radiation parameters E_{or}^a and γ_{or}^a , and in terms of the ground state material parameters P_o , K_o , γ_o , P_o^a , K_o^a , and γ_o^a .

6. MECHANICAL WAVES IN A SOLID. The temperature independent part of the nonrelativistic energy density for mechanical waves in a solid is given by¹⁰

$$\epsilon_{or}^a = \frac{1}{4} K_o^a k_a^2 A_a^2 \quad (91)$$

where

k_a = nonrelativistic wave number

A_a = nonrelativistic wave amplitude

The nonrelativistic Grüneisen parameter (diffuse radiation factor) for the radiation itself is given by¹⁰

$$\gamma_{or}^a = \frac{1}{3} + \frac{n}{W_a} \frac{dW_a}{dn} = \frac{1}{3} - \frac{n}{k_a} \frac{dk_a}{dn} \quad (92)$$

where $W_a = \omega/k_a$ = nonrelativistic phase velocity of the waves, and $\omega = 2\pi f$ where f = frequency of the waves. The corresponding relativistic expressions are written

$$\epsilon_{or} = \frac{1}{4} K_o k^2 A^2 \quad (93)$$

$$\gamma_{or} = \frac{1}{3} + \frac{n}{W} \frac{dW}{dn} = \frac{1}{3} - \frac{n}{k} \frac{dk}{dn} \quad (94)$$

where

k = relativistic wave number

A = relativistic wave amplitude

$W = \omega/k$ = relativistic phase velocity

As before the relativistic bulk modulus K_o is obtained from a solution of the ground state material equations (52) and (53).

The solution of the coupled equations (80) and (81) for the case of waves in a solid, described by equations (91) through (94), determines the relativistic radiation parameters k and A in terms of the nonrelativistic radiation parameters k_a and A_a and in terms of the material parameters in the following general forms

$$k = k(k_a, A_a, P_o, K_o, \gamma_o, P_o^a, K_o^a, \gamma_o^a, \omega) \quad (95)$$

$$A = A(k_a, A_a, P_o, K_o, \gamma_o, P_o^a, K_o^a, \gamma_o^a, \omega) \quad (96)$$

The relativistic phase velocity is $W = \omega/k$, so that it also is given by a function of the form

$$W = W(k_a, A_a, P_o, K_o, \gamma_o, P_o^a, K_o^a, \gamma_o^a, \omega) \quad (97)$$

The coupled relativistic wave equations (80) and (81) are complicated differential equations that are difficult to solve analytically, even for the simple case of a solid. However, some qualitative results can be obtained for the case of wave propagation in a solid by combining equations (80) and (93) and assuming that k and A (and hence W) are not density dependent. Actually, there is a density dependence of the phase velocity.⁹ But within the crude approximation that k and A are not density dependent, equations (80) and (93) yield

$$\frac{1}{4} k^2 A^2 \left(\ln \frac{dK_o}{dn} + JK_o \right) + g = \frac{1}{4} K_o^a k_a^2 A_a^2 \quad (98)$$

Then assuming $\gamma_{or} = 1/3$ and $\gamma_o = 1$ in equation (82) and (83) gives $I = 1$ and $J = -1$; and taking $K_o \sim n^\sigma$, where σ = adiabatic index, and finally $g \sim 2E_{or}$, gives the following approximate result for equation (98)

$$k^2 A^2 \sim \frac{K_o^a k_a^2 A_a^2}{(\sigma + 1)K_o} \quad (99)$$

But the solution of the ground state equations (52) and (53) is known to give $K_o \sim K_o^a$ at low pressures and $K_o \ll K_o^a$ at high pressures.³ Therefore within the limits of the approximations made to obtain equation (99), it follows that $kA \sim k_a A_a / \sqrt{\sigma + 1}$ for very low pressures, $kA \sim k_a A_a$ for moderate pressures, and $kA \gg k_a A_a$ for high pressures. The values of σ can range from zero to about two depending on the density of the system.

Finally, an explicit expression is given for the relativistic phase velocity of mechanical waves in a thermodynamic solid. The source terms for equations (80) and (81) are E_{or}^a and γ_{or}^a , where E_{or}^a is given by equation (91) while γ_{or}^a is obtained from equation (92) with³

$$\left(\frac{W_a}{c} \right)^2 = \frac{K_o^a}{E_o^a + P_o^a + K_o^a} \quad (100)$$

where c = light speed. The simultaneous solution of (80) and (81) gives E_{or} and γ_{or} . Then equation (94) is integrated to obtain the relativistic sound speed as follows

$$\frac{W}{c} = \exp \left[- \int_n^{\infty} \left(\gamma_{or} - \frac{1}{3} \right) \frac{dn}{n} \right] \quad (101)$$

Equation (94) shows that if W is an increasing function of density it follows that $\gamma_{or} > 1/3$ for mechanical waves, and therefore $W \leq c$ which is required by special relativity.

7. ELECTROMAGNETIC WAVES IN MATTER. An analogous calculation can be done for the case of the propagation of electromagnetic waves in matter. For this case the radiation energy density and pressure can also be written in the form of equations (57) through (60) except that now

$$E_{or} = \frac{1}{2} \left(\epsilon E^2 + \mu H^2 \right) \quad (102)$$

$$E_{or}^a = \frac{1}{2} \left(\epsilon_a E_a^2 + \mu_a H_a^2 \right) \quad (103)$$

where E and H = relativistic electric and magnetic radiation fields respectively; E_a and H_a = nonrelativistic electric and magnetic radiation fields respectively; ϵ , μ and ϵ_a , μ_a = relativistic and nonrelativistic permittivities and permeabilities respectively. Therefore the energy density is a function of many variables. The electroweak case includes even more parameters. Therefore it is not possible to obtain relativistic values of each parameter by solving the two coupled relativistic wave equations (80) and (81) unless the assumption is made that the electric and magnetic fields are unaffected by equations (80) and (81), and only the permittivity and permeability need to be considered. The solution of these two equations gives only the relativistic energy density and Grüneisen parameter for electromagnetic radiation. However an explicit solution for the relativistic phase velocity can be obtained.

The fundamental waves equations (80) and (81) determine E_{or} and γ_{or} in terms of E_{or}^a and γ_{or}^a , where E_{or}^a is given by equation (103) and γ_{or}^a is given by equation (92) with¹¹

$$W_a^2 = \frac{1}{\epsilon_a \mu_a} \quad (104)$$

Equation (94) is then integrated to obtain the relativistic phase velocity of electromagnetic waves in a thermodynamic medium as follows

$$\frac{W}{c} = \exp \left[- \int_0^n \left(\frac{1}{3} - \gamma_{or} \right) \frac{dn}{n} \right] \quad (105)$$

Equation (94) shows that if W is a decreasing function of density (as is the case for electromagnetic waves in a solid or low temperature quantum system) it follows that $\gamma_{or} < 1/3$ and $W \leq c$ for electromagnetic waves in matter. Note that the relativistic phase velocity depends on the material parameters P_0 , K_0 , γ_0 , P_0^a , K_0^a and γ_0^a .

8. CONCLUSION. Scale invariance has been demonstrated for a theory of relativistic thermodynamics that is based on a trace equation. The Grüneisen parameter must be introduced in addition to the pressure to insure local scale invariance. For solids and low temperature quantum systems, this means that the zero-temperature values of the pressure and Grüneisen parameter must be determined simultaneously through the solution of two coupled second order differential equations. The equations governing small amplitude waves in thermodynamic media can be obtained by a perturbation calculation on these ground state equations. Two coupled first order differential equations have been derived that describe relativistic wave propagation in solids and low temperature quantum systems. The solution of these wave equations determines the relativistic energy density and Grüneisen parameter for radiation in terms of the corresponding nonrelativistic radiation energy density and Grüneisen parameter and in terms of material parameters of the ground state. For mechanical waves the solution of the wave equations determines the relativistic amplitude and phase velocity.

Possible observable effects may occur in a number of high density systems. For instance, unusual dispersion effects and anomalously large wave amplitudes may be observed in the high pressure states of solids and liquids that occur in the interior of planets, stars, and stellar compact objects.^{12,13} Measurable effects may also occur in the vibrations of atomic nuclei that are associated with the giant nuclear resonances.¹⁴ Practical effects may also be noticed in the effects of nuclear explosions and in the interaction of high energy laser beams with solids.¹⁵ Finally, it should be pointed out that because the phase velocity and the wave amplitude must be calculated simultaneously from a pair of coupled nonlinear equations, it follows that the phase velocity depends on the wave amplitude, and this characteristic of nonlinear wave propagation may lead to the formation of shocks and solitons under certain conditions.¹⁶

APPENDIX A - RADIATION CONDITIONS. For radiation, the pressure is linearly related to the energy density as follows

$$P_r = \Gamma_r E_r \quad (A1)$$

where Γ_r = diffuse radiation factor given by¹⁰

$$\Gamma_r = \frac{1}{3} + \frac{n}{W} \frac{dW}{dn} \quad (A2)$$

where the phase velocity W is generally density dependent. Combining equation (A1) with equations (59) and (60) gives

$$P_{or} = \Gamma_r E_{or} \quad (A3)$$

$$P_{jr} = \Gamma_r E_{jr} \quad (A4)$$

But from the definition of the radiation Grüneisen parameter given by equation (64), it follows from equation (A4) that

$$\Gamma_r = \gamma_{or} \quad (A5)$$

which means that the radiation Grüneisen parameter is equal to the diffuse radiation factor.

If the pressure and the internal energy of the radiation are written as

$$P_r = P_{or} + P_{jr} T^j \quad (A6)$$

$$U_r = V E_r = U_{or} + U_{jr} T^j \quad (A7)$$

then the application of the Gibbs-Helmholtz equation of thermodynamics

$$\left(\frac{\partial U_r}{\partial V} \right)_T = T \left(\frac{\partial P_r}{\partial T} \right)_V - P_r \quad (A8)$$

yields the following completely general equations

$$P_{or} = - \frac{dU_{or}}{dV} = - \frac{d}{dV} (VE_{or}) \quad (A9)$$

$$(j-1)P_{jr} = \frac{dU_{jr}}{dV} = \frac{d}{dV} (VE_{jr}) \quad (A10)$$

Combining equations (A3) and (A4) with equations (A9) and (A10) respectively gives

$$\frac{d}{dV} (VE_{or}) + \gamma_{or} E_{or} = 0 \quad (A11)$$

$$\frac{d}{dV} (VE_{jr}) - (j-1)\gamma_{or} E_{jr} = 0 \quad (A12)$$

or equivalently

$$\frac{V}{U_{or}} \frac{dU_{or}}{dV} = - \gamma_{or} \quad (A13)$$

$$\frac{V}{U_{jr}} \frac{dU_{jr}}{dV} = (j-1)\gamma_{or} \quad (A14)$$

The radiation equations (A13) and (A14) can be immediately integrated to give

$$U_{or} = D_{or} \exp \left(\int^n \gamma_{or} \frac{dn}{n} \right) \quad (A15)$$

$$U_{jr} = D_{jr} \exp \left[- (j-1) \int^n \gamma_{or} \frac{dn}{n} \right] \quad (A16)$$

and finally,

$$E_{or} = nD_{or} \exp \left(\int^n \gamma_{or} \frac{dn}{n} \right) \quad (A17)$$

$$E_{jr} = n D_{jr} \exp \left[- (j-1) \int^n \gamma_{or} \frac{dn}{n} \right] \quad (A18)$$

Then combining equation (A18) with equation (54) gives

$$\frac{E_{jr}}{E_j} = \frac{D_{jr}}{C_j} \exp \left[- (j-1) \int^n (\gamma_{or} - \gamma_o) \frac{dn}{n} \right] \quad (A19)$$

which is the desired result.

APPENDIX B - EVALUATION OF $d\delta_{or}/dn$. The derivative that occurs in equation (72) can be evaluated from equation (63) as follows,

$$n \frac{d\delta_{or}}{dn} = (\gamma_{or} - \gamma_o) n \frac{d}{dn} \left(\frac{E_{jr}}{E_j} \right) + \frac{E_{jr}}{E_j} \left(n \frac{d\gamma_{or}}{dn} - n \frac{d\gamma_o}{dn} \right) \quad (B1)$$

In order to evaluate the derivative of the energy ratio term in equation (B1) one uses equation (A19) of Appendix A to get the following result

$$n \frac{d}{dn} \left(\frac{E_{jr}}{E_j} \right) = - (j-1) \frac{E_{jr}}{E_j} (\gamma_{or} - \gamma_o) \quad (B2)$$

Combining equations (B1) and (B2) gives the result

$$n \frac{d\delta_{or}}{dn} = \frac{E_{jr}}{E_j} \left[n \frac{d\gamma_{or}}{dn} - n \frac{d\gamma_o}{dn} - (j-1) (\gamma_{or} - \gamma_o)^2 \right] \quad (B3)$$

which is the desired result.

REFERENCES

1. Aitchinson, I. and Hey, A., Gauge Theories in Particle Physics, Adam Hilger Ltd., Bristol, United Kingdom, 1982.
2. Leader, E. and Predazzi, E., Gauge Theories and the 'New Physics', Cambridge University Press, London, p. 29, 1982.

3. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.
4. Aitchinson, I., An Informal Introduction to Gauge Field Theories, Cambridge University Press, p. 155, 1982.
5. Zharkov, V. N. and Kalinin, V. A., Equations of State for Solids at High Pressures and Temperatures, Consultants Bureau, New York, 1971.
6. Varley, J., "The Thermal Expansion of Pure Metals and the Possibility of Negative Coefficients of Volume Expansion," Proc. R. Soc. A237, 413, 1956.
7. Barron, T., Collins, J. and White, G., "Thermal Expansion of Solids at Low Temperatures," Adv. in Physics, 29, 609, 1980.
8. Grot, R. A. and Eringen, A. C., "Relativistic Continuum Mechanics, Part I - Mechanics and Thermodynamics; and Part II - Electromagnetic Interactions with Matter," Int. J. Engng. Sci. Vol 4, pp. 611-670, Pergamon Press, New York, 1966.
9. Tolstoy, I., "On Elastic Waves in Prestressed Solids," Journal of Geophysical Research, Vol 87, No. B8, pp. 6823-6827, Aug 10 1982.
10. Brillouin, L., Tensors in Mechanics and Elasticity, Academic Press, New York, 1964.
11. Born, M. and Wolf, E., Principles of Optics, Pergamon Press, New York, 1959.
12. Zeldovich, Ya. B. and Novikov, I. D., Relativistic Astrophysics, Vol. I Stars and Relativity, University of Chicago Press, 1978.
13. Misner, C. W., Thorne, K. S. and Wheeler, J. A., Gravitation, W. H. Freeman, San Francisco, 1973.
14. Vazquez, A., "Giant Resonances as Oscillations of Two Elastically Coupled Fluids," Phys. Rev. Lett., Vol 50, No. 22, pp. 1756-1758, 30 May 1983.
15. Rodean, H. C., Nuclear-Explosion Seismology, U. S. Atomic Energy Commission, AEC Critical Review Series, 1971.
16. Dodd, R. K., Eilbeck, J. C., Gibbon, J. D. and Morris, H. C., Solitons and Nonlinear Wave Equations, Academic Press, New York, 1982.

Regularity Results for the Porous Medium Equation

Klaus Höllig

Computer Sciences Department and
Mathematics Research Center
University of Wisconsin-Madison
Madison, Wisconsin 53706

and

Heinz-Otto Kreiss

Department of Applied Mathematics
California Institute of Technology
Pasadena, California 91125

ABSTRACT. The equation $u_t = \Delta u^m$ models the expansion of a gas with density $u(x, t)$ in a porous medium. We give an equivalent formulation of this equation as a free boundary problem, the free surface being the boundary of the set where the gas density is nonzero. Using this formulation, we discuss new a priori estimates for the pressure $v := (m/(m-1))u^{m-1}$. In one dimension, our estimates imply that the pressure and the free boundary are infinitely differentiable.

1. THE FREE BOUNDARY PROBLEM. The initial value problem

$$\begin{aligned}u_t(x, t) &= \Delta u^m(x, t), \quad x \in \mathbf{R}^d, \quad t \leq T, \\u(x, 0) &= u_0(x)\end{aligned}\tag{1}$$

describes the expansion of a gas in a porous medium: u denotes the gas density, $\Delta = \nabla \cdot \nabla = \partial_1^2 + \dots + \partial_d^2$ and $m > 1$ is a physical parameter.

In this section we derive an equivalent formulation of the initial value problem (1). We assume that u_0 has bounded support $\Omega(0)$ with smooth boundary $\partial\Omega(0)$. Then, $\Omega(t) := \text{supp } u(\cdot, t)$ is bounded for all t . We rewrite (1) as an initial value problem for the pressure $v := (m/(m-1))u^{m-1}$. By direct substitution,

$$\begin{aligned}v_t(x, t) &= (m-1)v(x, t)\Delta v(x, t) + |\nabla v(x, t)|^2, \quad x \in \Omega(t), \quad t \leq T, \\v(x, 0) &= v_0(x).\end{aligned}\tag{2}$$

We assume that v is twice continuously differentiable and define a family of curves $y \rightarrow \xi(y, t)$, $y \in \Omega(0)$, via the system

$$\begin{aligned}\xi_t(y, t) &= -\nabla v(\xi(y, t), t) \\ \xi(y, 0) &= y.\end{aligned}\tag{3}$$

Note that

$$\frac{d}{dt}v(\xi(y, t), t) = (m-1)v(\xi(y, t), t)\Delta v(\xi(y, t), t)$$

which implies in particular that $v(\xi(y, t), t) = 0$ for $y \in \partial\Omega(0)$. Therefore, if $t \leq T$ with T sufficiently small, $\xi(\cdot, t)$ defines a 1-1 mapping of $\Omega(0)$ onto $\Omega(t)$.

Summarizing the above considerations we have:

If u is a weak solution of (1) for which v is twice continuously differentiable on its (closed) support, then v is a classical solution of (2) and, for T sufficiently small, $\Omega(t) = \xi(\Omega(0), t)$ with ξ defined by (3).

The converse of this statement also holds:

If the pair (v, ξ) is a classical solution of (2,3) with $\Omega(t) = \xi(\Omega(0), t)$ and if the mapping $\xi(\cdot, t)$ is 1-1 for all $t \leq T$, then u is a weak solution of (1).

Here we have used the notation $f(I) := \{f(\lambda) : \lambda \in I\}$ for a function f and a set I .

To verify the last assertion, let ϕ be any test function with compact support in $Q := \mathbf{R}^d \times (0, T)$. Integrating by parts,

$$\begin{aligned} \int_Q u \phi_t - u^m \Delta \phi &= \int_0^T \int_{\Omega(t)} u \phi_t \\ &+ \int_Q (\Delta u^m) \phi - \int_0^T \int_{\partial\Omega(t)} (-\nabla v / |\nabla v|) \cdot (\nabla u^m) \phi, \end{aligned} \quad (4)$$

where we have used that v vanishes on the boundary of $\Omega(t)$ and therefore the boundary normal is parallel to $-\nabla v$.

We need the following

Lemma. Assume that $\xi(\cdot, t)$ as defined in (3) is a 1-1 mapping of $\Omega(0)$ onto $\Omega(t)$. Then, for any smooth function f ,

$$\int_{\Omega(t)} f_t = \partial_t \int_{\Omega(t)} f - \int_{\partial\Omega(t)} |\nabla v| f. \quad (5)$$

Using this and the fact that $\phi(x, \cdot)$ has compact support in $(0, T)$, the right hand side of (4) equals

$$\begin{aligned} - \int_0^T \int_{\Omega(t)} u_t \phi &- \int_0^T \int_{\partial\Omega(t)} |\nabla v| u \phi \\ &+ \int_Q (\Delta u^m) \phi - \int_0^T \int_{\partial\Omega(t)} (-\nabla v / |\nabla v|) \cdot (\nabla u^m) \phi. \end{aligned}$$

This expression vanishes since equations (1) and (2) are equivalent on $\Omega(t)$ and since $\nabla u^m = u \nabla v$.

For the sake of completeness we include a

Proof of the Lemma. Denote by $\nabla \xi(\cdot, t)$ the Jacobi matrix $\{\partial_k \xi_j(\cdot, t)\}$ where ξ_j is the j -th component of the vector ξ . From (3) it follows that the determinant of $\nabla \xi$ satisfies the differential equation

$$(\det \nabla \xi)_t = -\Delta v (\det \nabla \xi).$$

Note in particular that $\det \nabla \xi(\cdot, t)$ is positive for all t . Using this, we obtain

$$\begin{aligned}
 \int_{\Omega(t)} f_t(x, t) \, dx &= \int_{\Omega(0)} f_t(\xi(y, t), t) (\det \nabla \xi)(y, t) \, dy \\
 &= \partial_t \int_{\Omega(0)} f(\xi(y, t), t) (\det \nabla \xi)(y, t) \, dy \\
 &\quad - \int_{\Omega(0)} (\nabla f)(\xi(y, t), t) \cdot \xi_t(y, t) (\det \nabla \xi)(y, t) \, dy \\
 &\quad + \int_{\Omega(0)} f(\xi(y, t), t) \Delta v(\xi(y, t), t) (\det \nabla \xi)(y, t) \, dy \\
 &= \partial_t \int_{\Omega(t)} f(x, t) \, dx + \int_{\Omega(t)} \nabla f(x, t) \cdot \nabla v(x, t) + f(x, t) \Delta v(x, t) \, dx.
 \end{aligned} \tag{6}$$

Equation (5) follows by replacing the last term in (6) by a boundary integral and recalling that the boundary normal equals $-\nabla v / |\nabla v|$.

One notices the similarity of the free boundary problem (2,3) to the Stefan problem. E.g. the one phase Stefan problem is described by equation (3) and the heat equation

$$\begin{aligned}
 v_t(x, t) &= \Delta v(x, t), \quad x \in \Omega(t), \quad t \leq T, \\
 v(x, 0) &= v_0(x).
 \end{aligned}$$

In this case, v denotes the temperature distribution of water in a region $\Omega(t)$ which is surrounded by ice.

Existence of weak solutions for the porous medium equation and Stefan problems can be proved via semi group theory [3]. Regularity results, in particular for the free boundaries, appear to be considerably more difficult to obtain. For the one phase Stefan problem the existence and C^∞ -regularity of classical solutions was obtained by Hanzawa [7] and Kinderlehrer and Nirenberg [11] using rather sophisticated techniques. For the porous medium equation in several variables no comparable regularity results are known. A major difficulty in studying existence and regularity for the system (2,3) is that the parabolic operator degenerates on the free surface.

For the one dimensional case ($d = 1$) Caffarelli and Friedman [4] have shown the existence of classical solutions for the system (2,3). Using their result we proved in [9] that the pressure v and the free boundary are infinitely differentiable as soon as the support of $u(\cdot, t)$ is expanding. This result is discussed in section 2. In section 3, we describe new a priori estimates for the multivariate case. If a suitable regularization or approximation procedure can be found, these estimates imply the existence of smooth solutions for the system (2,3).

2. C^∞ -REGULARITY IN ONE DIMENSION. In one variable we may assume that $\Omega(0) = [-1, 1]$. Then, the system (2,3) reduces to

$$\begin{aligned}\xi'(y, t) &= -v_x(\xi(y, t), t), \quad y = \pm 1, \\ \xi(\pm 1, 0) &= \pm 1\end{aligned}\quad (7)$$

$$\begin{aligned}v_t &= (m-1)vv_{xx} + v_x^2, \quad \xi(-1, t) \leq x \leq \xi(1, t), \quad t > 0, \\ v(x, 0) &= v_0.\end{aligned}\quad (8)$$

The curves $\xi(\pm 1, \cdot)$ are Lipschitz continuous [12], but in general not C^1 . As was observed in [1], ξ' need not be continuous at

$$t_y := \sup\{t : \xi(y, t) = y\}, \quad y = \pm 1.$$

Caffarelli and Friedman [4] proved that the system (7,8) has a classical solution for $t > \max\{t_{-1}, t_1\}$, i.e. the functions v, v_t, v_x and vv_{xx} are continuous up to the free boundaries. Using this result we proved in [9] the following optimal regularity result.

Theorem. Assume that v_0 is continuous. Then, $\xi(y, \cdot) \in C^\infty(t_y, \infty)$, $y = \pm 1$, and $v \in C^\infty(\Omega_+)$ where $\Omega_+ := \{(x, t) : v(x, t) > 0\} \cup \{(\xi(y, t), t) : t > t_y, y = \pm 1\}$.

Our proof of this result is based on a priori estimates in weighted norms. To illustrate the main idea, we make a simplifying assumption. We suppose that $t_{-1} = t_1 = 0$ which implies in particular that

$$v_x(\xi(y, t), t) \neq 0, \quad y = \pm 1, \quad t \geq 0.$$

Under this strengthened hypothesis we have the following a priori estimate.

Proposition. For any smooth solution (v, ξ) of (7,8) and any $k > 0$, $\delta > 0$,

$$\max_{\delta \leq t \leq T} \int_{\Omega(t)} v |\partial_x^k v|^2 dx + \int_\delta^T \int_{\Omega(t)} v^2 |\partial_x^{k+1} v|^2 dx dt \leq A_k \quad (9)$$

where the constant A_k depends on k, δ, T and v_0 but not on v .

An analogous estimate is also valid for a suitably defined Galerkin approximation. This yields the regularity of v and, by (7), also the regularity of ξ .

We prove (9) by induction on k . Thus assume that the estimate (9) holds for $k < l$. Differentiating (8) and setting $w := \partial_x^l v$ we obtain

$$w_t = (m-1)vw_{xx} + [2 - l(m-1)]v_x w_x + f \quad (10)$$

where f is a sum of terms of the form

$$c_{\nu\mu} \partial_x^\nu v \partial_x^\mu v, \quad \nu + \mu = l + 2, \quad \nu \leq \mu \leq l.$$

We multiply (10) by $\tau^2 vw$ with $\tau := t - \delta$ and integrate each term over the interval $\Omega(t) = [\xi(-1, t), \xi(1, t)]$. Integrating by parts and observing that the boundary terms vanish we obtain

$$\begin{aligned} \int_{\Omega(t)} \tau^2 vw w_t &= \frac{1}{2} \partial_t \int_{\Omega(t)} \tau^2 v w^2 \\ &- \int_{\Omega(t)} \tau v w^2 - \frac{1}{2} \int_{\Omega(t)} \tau^2 (m-1) v v_{xx} w^2 - \frac{1}{2} \int_{\Omega(t)} \tau^2 v_x^2 w^2 \\ &=: \frac{1}{2} \partial_t \int \tau^2 v w^2 - I_1 - ((m-1)/2) I_2 - (1/2) I_3 \end{aligned}$$

$$\begin{aligned} (m-1) \int_{\Omega(t)} \tau^2 v^2 w w_{xx} &= \\ &= (m-1) \int \tau^2 v^2 w_x^2 - 2(m-1) \int \tau^2 v v_x w w_x \end{aligned}$$

$$\begin{aligned} [2 + (l-2)(m-1)] \int_{\Omega(t)} \tau^2 v v_x w w_x &= \\ &= [1 + (l-2)(m-1)/2] (I_2 - I_3). \end{aligned}$$

Combining these identities yields

$$\begin{aligned} \frac{1}{2} \partial_t \int_{\Omega(t)} \tau^2 v w^2 + (m-1) \int_{\Omega(t)} \tau^2 v^2 w_x^2 &= \\ &+ I_1 - \gamma_2 I_2 - \gamma_3 I_3 - \int \tau^2 v w f \end{aligned} \quad (11)$$

where the constant γ_3 is positive for $l > 1$.

The estimation of the terms on the right hand side of (11) is somewhat technical, in particular if l is small. Let us assume that $l > 4$ and refer to [9] for the remaining cases. We need two auxiliary inequalities. By the result in [4], v is continuously differentiable with modulus of continuity depending only on v_0 . Using this, and the fact that v_x does not vanish on either of the free boundaries, it is not difficult to show (cf. Lemmas 1, 2 in [9]) that

$$\|g^2\|_{\infty, \Omega(t)} \leq c_1 \int_{\Omega(t)} (g^2 + g_x^2) \leq c_2 \int_{\Omega(t)} v(g^2 + g_{xx}^2) \quad (12)$$

for any smooth function g , uniformly for $t \leq T$. Moreover, there exists a positive constant c_3 which depends only on v_0 and T such that

$$v(x, t)^2 + v_x(x, t)^2 > c_3, \quad x \in \Omega(t). \quad (13)$$

Using Hölder's inequality and (13) yields

$$\begin{aligned} I_1 &\leq \epsilon \int \tau^2 w^2 + \epsilon^{-1} \int v^2 w^2 \\ &\leq (\epsilon/c_3) \int \tau^2 v_x^2 w^2 + (\epsilon^{-1} + \epsilon T^2/c_3) \int v^2 w^2. \end{aligned} \quad (14)$$

From (12) it follows that $|v_{xx}(\cdot, t)|_{\infty, \Omega(t)}$ can be bounded in terms of A_4 for $t \geq \delta$. Therefore, if K denotes a generic constant which may depend on A_{l-1} and v_0 , we have

$$|I_2| \leq K \int \tau^2 v w^2. \quad (15)$$

For the last term in (11) we have to estimate the integrals $I_{\nu\mu} := \int \tau^2 v (\partial_x^\nu v) (\partial_x^\mu v) w$. Since $\nu \leq (l+2)/2 \leq l-3$, (12) implies that $|\partial_x^\nu v(\cdot, t)|_{\infty, \Omega(t)} \leq K$ for $t \geq \delta$. Hence we obtain

$$|I_{\nu\mu}| \leq K(1 + \int \tau^2 v w^2). \quad (16)$$

We choose ϵ/c_3 less than γ_3 so that the first term on the right hand side of (14) is less than $\gamma_3 I_3$. Substituting the estimates (14-16) into (11) we obtain the differential inequality

$$\begin{aligned} \frac{1}{2} \partial_t \int_{\Omega(t)} \tau^2 v w^2 + (m-1) \int_{\Omega(t)} \tau^2 v^2 w_x^2 \\ \leq K(1 + \int \tau^2 v w^2) + (\epsilon^{-1} + \epsilon T^2/c_3) \int v^2 w^2. \end{aligned}$$

Integrating this inequality over the interval $[\delta, T]$ and using that

$$\int_\delta^T \int v^2 w^2 \leq A_{l-1}$$

yields (9) for $k = l$ with δ replaced by 2δ .

3. A PRIORI ESTIMATES IN SEVERAL VARIABLES. Based on the results in the univariate case one is tempted to conjecture that the free surface $\partial\Omega(t)$ is infinitely differentiable in the neighborhood of any point where the support of $v(\cdot, t)$ is expanding. A first step in this direction is again a result by Caffarelli and Friedman [5] who showed that the free surface is Hölder continuous. More recently, Gurtin, McCamy and Socolovsky [6] discovered an interesting transformation which yields a system of partial differential equations for the free surface which does not involve v . However, this system does not fall in any of the standard categories and, except for the univariate case [8] existence of smooth solutions could not been established. We formulate below a priori estimates which would imply the existence of classical solutions for the system (2,3) if an appropriate approximation procedure can be found.

Theorem. Assume that $\partial\Omega(0)$ is smooth, $\nabla v_0(x) \neq 0$ for $x \in \partial\Omega(0)$ and (v, ξ) is a smooth solution of the system (2,3). Let $|w|_{k,\Omega}^2 := \sum_{|\alpha|=k} \int_{\Omega} |\partial^\alpha w|^2$ denote the semi norm of $H^k(\Omega)$. Then, for any $k > 2 + d/2$, there exist constants T and B which depend on k , $\Omega(0)$ and v_0 but not on v and ξ such that

$$\max_{0 \leq t \leq T} |v(\cdot, t)|_{k,\Omega(t)} \leq B.$$

The proof of the Theorem is given in a forthcoming paper [10]. As in the univariate case, our arguments are based on energy estimates. The major difficulty is the estimation of the tangential derivatives of v in a neighborhood of $\partial\Omega(t)$.

References

- [1] D. G. Aronson, Regularity properties of flows through porous media: A counterexample, *SIAM J. Appl. Math.* **19** (1970), 299-307.
- [2] D. G. Aronson, L. A. Caffarelli, and J. L. Vazquez, Interfaces with a corner point in one-dimensional porous-medium flow, *Lefschetz Center for Dynamical Systems*, report #84-9.
- [3] P. Benilan, M.G. Crandall and A. Pazy, *M-accretive operators* (in preparation).
- [4] L. Caffarelli and A. Friedman, Regularity of the free boundary for the one-dimensional flow of gas in a porous medium. *Amer. J. Math.* **101** (1979), 1193-1218.
- [5] L. Caffarelli and Avner Friedman, Regularity of the free boundary in an n -dimensional porous medium, *Indiana Univ. Math. J.* **29** (1980), 361-391.
- [6] M. Gurtin, R. MacCamy and E. Socolovsky, A coordinate transformation for the porous media equation that renders the free boundary stationary, to appear.
- [7] E. Hanzawa, Classical solutions of the Stefan problem, *Tohoku Math. J.* **33** (1981), 297-335.
- [8] K. Höllig and M. Pilant, Regularity of the free boundary for the porous medium equation, to appear in *Indiana Univ. Math. J.*
- [9] K. Höllig and H.-O. Kreiss, C^∞ -regularity for the porous medium equation, submitted to *Math. Zeitschrift*.
- [10] K. Höllig and H.-O. Kreiss, A priori estimates for the porous medium equation in several variables, ms.
- [11] D. Kinderlehrer and L. Nirenberg, The smoothness of the free boundary in the one phase stefan problem. *Comm. Pure Appl. Math.* **31** (1978), 257-282.
- [12] B. F. Knerr, The porous medium equation in one-dimension, *Trans. Amer. Math. Soc.* **234** (1977), 381-415.

ON THE TREATMENT OF POISSON'S EQUATION BY PIECEWISE
POLYNOMIALS AND PARTITION METHOD

SHIH C. CHU

TECHNOLOGY BRANCH, ARMAMENT DIVISION
FIRE CONTROL & SMALL CALIBER WEAPON SYSTEMS LABORATORY
US ARMY ARMAMENT RESEARCH AND DEVELOPMENT CENTER
DOVER, NEW JERSEY 07801-5001

ABSTRACT

An efficient numerical method, used previously for a ordinary differential equation is here extended to a partial differential equation (in particular, Poisson's equation) of a function of two variables. Bicubic functions are used as the basic approximations. Residuals are liquidated by setting their integrals equal to zero over specified subregions of analyticity.

INTRODUCTION

The approximate solution of ordinary differential equations with the use of piecewise polynomials, or spline curves, and the partition method was investigated by Langhaar and Chu (1). From the results of their application of piecewise cubic and piecewise quintic approximations to a number of ordinary differential equations, they have concluded that this method has certain advantages over other methods. For example, this method yields directly the first derivatives in the case of a piecewise cubic, or the first and second derivatives in the case of a piecewise quintic approximation in addition to the value of the function at a finite number of points. The finite-difference method of solution yields directly only the values of the function, an obvious drawback, since for many problems in mechanics, the derivatives are the sought for unknowns. Also, the finite-element method is based on variational principles and the physical problem at hand. The piecewise polynomial method is free of such formulations, and is, in that sense, more general.

In this investigation, the piecewise polynomial method is extended to partial differential equations (in particular, Poisson's equation) of a function of two variables.

For an approximating polynomial, a bicubic was chosen, i.e.,

$$\tilde{\phi}(x,y) = \sum_{i=0}^3 \sum_{j=0}^3 \alpha_{ij} x^i y^j \quad (1)$$

CONSTRUCTION OF THE BICUBIC POLYNOMIALS

A finite closed region of the x-y plane is divided into MN subregions (M being the number of subregions in the x-direction and N being the number of subregions in the y-direction) by the lines $x = x_i$ and $y = y_j$ such that

$$0 = x_0 < x_1 < x_2 \dots < x_m = L$$

$$0 = y_0 < y_1 < y_2 \dots < y_n = W$$

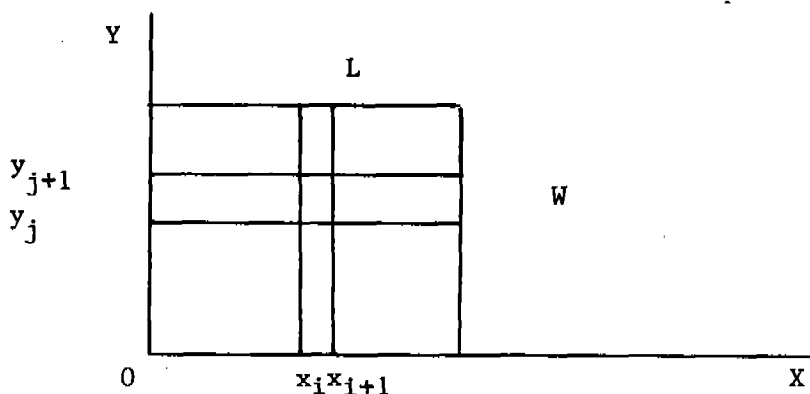


Figure 1

Equation 1 is written in the following explicit form

$$\begin{aligned} \tilde{\phi}(x,y) = & a_1 + a_2x + a_3x^2 + a_4x^3 \\ & + a_5y + a_6xy + a_7x^2y^2 + a_8x^3y^2 \\ & + a_9y^2 + a_{10}xy^2 + a_{11}x^2y^2 + a_{12}x^3y^2 \\ & + a_{13}y^3 + a_{14}xy^3 + a_{15}x^2y^3 + a_{16}x^3y^3 \end{aligned}$$

In terms of values ϕ , ϕ_x , ϕ_y and ϕ_{xy} at the four corners of the subregion, one has

$$\begin{aligned}\tilde{\phi}(x_i, y_j) &= a_1 + a_2 x_i + a_3 x_i^2 + a_4 x_i^3 + \dots + a_{13} y_j^3 + a_{14} x_i y_j^3 \\ &\quad + a_{15} x_i^2 y_j^3 + a_{16} x_i^3 y_j^3 = \phi_1\end{aligned}$$

$$\tilde{\phi}_x(x_i, y_j) = a_2 + 2a_3 x_i + 3a_4 x_i^2 + \dots + a_{14} y_j^3 + 2a_{15} x_i y_j^3 + 3a_{16} x_i^2 y_j^3 = \phi_2$$

$$\begin{aligned}\tilde{\phi}_y(x_i, y_j) &= a_5 + a_6 x_i + a_7 x_i^2 + a_8 x_i^3 + \dots + 3a_{13} y_j^2 + 3a_{14} x_i y_j^2 + 3a_{15} x_i^2 y_j^2 \\ &\quad + 3a_{16} x_i^3 y_j^2 = \phi_3\end{aligned}$$

$$\tilde{\phi}_{xy}(x_i, y_j) = a_6 + 2a_7 x_i + 3a_8 x_i^2 + 2a_{10} y_j + \dots + 6a_{15} x_i y_j^2 + 9a_{16} x_i^2 y_j^2 = \phi_4$$

$$\phi(x_i, y_{j+1}) = a_1 + a_2 x_i + a_3 x_i^2 + \dots + a_{14} x_i y_{j+1}^3 + a_{15} x_i^2 y_{j+1}^3 + a_{16} x_i^3 y_{j+1}^3 = \phi_5$$

⋮

$$\begin{aligned}\phi(x_{i+1}, y_j) &= a_1 + a_2 x_{i+1} + a_3 x_{i+1}^2 + \dots + a_{14} x_{i+1} y_j^3 + a_{15} x_{i+1}^2 y_j^3 \\ &\quad + a_{16} x_{i+1}^3 y_j^3 = \phi_9\end{aligned}$$

⋮

$$\begin{aligned}\phi(x_{i+1}, y_{j+1}) &= a_1 + a_2 x_{i+1} + a_3 x_{i+1}^2 + \dots + a_{14} x_{i+1} y_{j+1}^3 + a_{15} x_{i+1}^2 y_{j+1}^3 \\ &\quad + a_{16} x_{i+1}^3 y_{j+1}^3 = \phi_{13}\end{aligned}$$

⋮

$$\begin{aligned}\phi_{xy}(x_{i+1}, y_{j+1}) &= a_6 + 2a_7 x_{i+1} + 3a_8 x_{i+1}^2 + \dots + 6a_{15} x_{i+1} y_{j+1}^2 + 9a_{16} x_{i+1}^2 y_{j+1}^2 \\ &= \phi_{16}\end{aligned}$$

These equations when written in the form of matrices, are

$$\begin{bmatrix} 1 & x_i & x_i^2 & . & . & . & x_i^2 y_j^3 & x_i^3 y_j^3 \\ 0 & 1 & 2x_i & . & . & . & 2x_i y_j^3 & 3x_i^2 y_j^3 \\ 0 & 0 & 0 & . & . & . & 3x_i^2 y_j^2 & 3x_i^3 y_j^2 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 3x_{i+1}^2 y_{j+1}^2 & 3x_{i+1}^3 y_{j+1}^2 \\ 0 & 0 & 0 & . & . & . & 6x_{i+1}^2 y_{j+1}^2 & 9x_{i+1}^3 y_{j+1}^2 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ . \\ a_{15} \\ a_{16} \end{Bmatrix} = \begin{Bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ . \\ . \\ . \\ \phi_{15} \\ \phi_{16} \end{Bmatrix}$$

or just

$$[E] [A] = [\Phi] \quad (2)$$

The bicubic polynomial, equation 1, when written in the form of matrices is

$$\phi(x,y) = [1,x,x^2,x^3, \dots, x^2 y^3, x^3 y^3] [A] \quad (3)$$

or alternatively

$$\phi(x,y) = [1,x,x^2,x^3, \dots, x^2 y^3, x^3 y^3] [C] [\Phi] \quad (4)$$

where [C] is as yet some undetermined 16 x 16 matrix of constants.

Comparison of equations 3 and 4 shows that

$$[A] = [C] [\Phi]$$

but, from equation 2

$$[A] = [E]^{-1} [\Phi]$$

Therefore, $[C] = [E]^{-1}$

and a bicubic polynomial whose coefficients are expressed in the generalized deflection ϕ_1 can now be written for each subregion.

POISSON'S EQUATION

Poisson's equation is

$$L(\phi) = f(x,y) \quad (5)$$

where

$$L \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

If $\tilde{\phi}$ is substituted for ϕ in equation 5, then

$$L(\tilde{\phi}) = [L(1), L(x), L(x^2), \dots, L(x^2y^3), L(x^3y^3)][C][\tilde{\phi}] = f(x,y) + e(x,y)$$

where $e(x,y)$ is the error resulting from the approximation of ϕ by $\tilde{\phi}$.

Each subregion is divided into four parts and the integral of the error $e(x,y)$ over each of these parts is set equal to zero (Partition Method). This procedure yields a set of $4MN$ equations involving the $4(M+1)(N+1)$ unknown ϕ_i , i.e.,

$$\begin{aligned} \iint_{\text{each part}} L(\tilde{\phi}) dx dy &= \iint_{\text{each part}} [L(1), L(x), L(x^2), \dots, L(x^2y^3), L(x^3y^3)] dx dy [C][\tilde{\phi}] \\ &= \iint_{\text{each part}} f(x,y) dx dy \end{aligned}$$

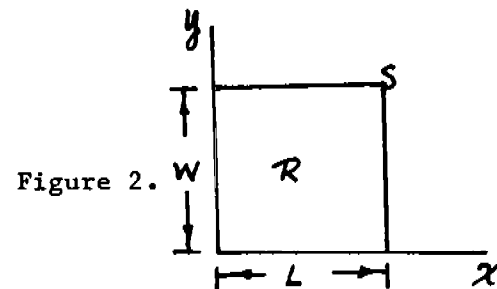
The remaining $4(M+N+1)$ equations needed are obtained from the boundary conditions.

EXAMPLE

Consider the torsion of a bar of rectangular cross-section, Fig. 2. This problem reduces to the solution of

$$\begin{aligned} \nabla^2 \phi &= 1 \text{ in } R \\ \phi &= 0 \text{ on } S \end{aligned}$$

where the shear stresses are given by



$$\tau_{xz} = -2G\theta \frac{\partial \phi}{\partial y}$$

$$\tau_{yz} = -2G\theta \frac{\partial \phi}{\partial x}$$

Let $L=W=1$ and $M=N=2$, then the unit square region R is divided into four equal subregions each of which is divided into four equal parts. Setting the integral of the error over each of these parts equal to zero yields sixteen equations.

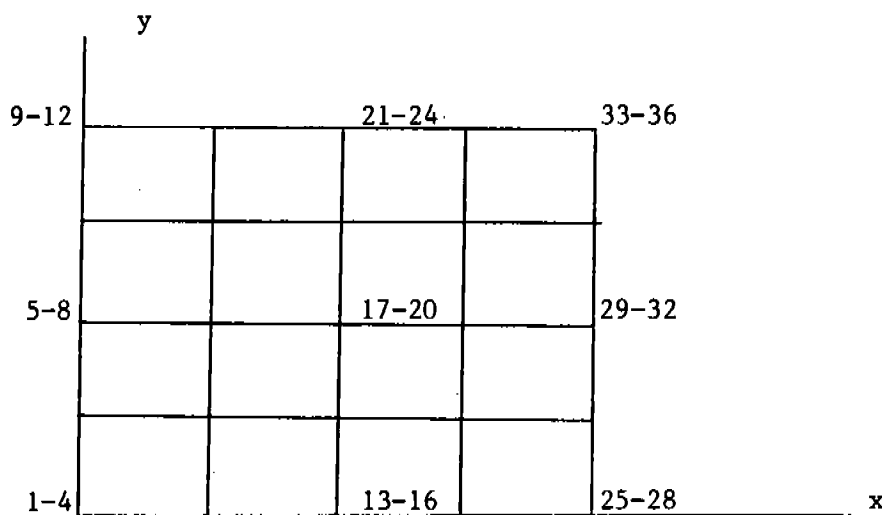


Figure 3.

The numbering scheme is shown in Figure 3. At the lower left node,

$$\phi_1 = \phi, \quad \phi_2 = \phi_x, \quad \phi_3 = \phi_y, \quad \phi_4 = \phi_{xy}, \text{ etc.}$$

Imposing the boundary condition ϕ on S yields

$$\phi_1 = \phi_5 = \phi_9 = \phi_{13} = \phi_{21} = \phi_{25} = \phi_{29} = \phi_{33} = 0$$

Also, since $\phi = \text{constant}$ along S

$$\phi_2 = \phi_{10} = \phi_{14} = \phi_{22} = \phi_{26} = \phi_{34} = 0$$

and

$$\phi_3 = \phi_7 = \phi_{11} = \phi_{27} = \phi_{31} = \phi_{35} = 0$$

The above 20 equations, together with the 16 equations obtained from setting the integral of the error equal to zero, constitute a set of 36 linear algebraic equations from which the 36 deflections ϕ_i can be obtained. Knowing the ϕ_i , the problem is completely solved.

Alternatively, the integral of the error along the boundary can be set equal to zero, along 16 equal distinct segments of the boundary S together with the 4 equations

$$\phi_1 = \phi_9 = \phi_{25} = \phi_{33} = 0$$

Nonzero values of ϕ , ϕ_x , ϕ_y , and ϕ_{xy} are given below:

$\phi_{xy} = 2.0357$	at (0, 0)
$\phi_x = 0.3304$	at (0, $1/2$)
$\phi_{xy} = -2.0357$	at (0, 1)
$\phi_y = 0.3304$	at ($1/2$, 0)
$\phi = 0.0737$	at ($1/2$, $1/2$)
$\phi_y = -0.3304$	at ($1/2$, 1)
$\phi_{xy} = -2.0357$	at (1, 0)
$\phi_x = -0.3304$	at (1, $1/2$)
$\phi_{xy} = 2.0357$	at (1, 1)

The results are shown in Table 1 for ϕ_x at the midpoint of the edge $x = 0$, i.e., ϕ in Figure 2. The value for ϕ is also given at the center of the region R, i.e., ϕ_{17} . These values are given for $M=N=2$ and $M=N=4$, along with the exact values. The values of ϕ_i found by satisfying the boundary conditions pointwise were the same as those found by setting the integral of the error along the boundary equal to zero.

$M = N$	$\phi_6 = \phi_x(0, 1/2)$	$\phi_{17} = \phi(1/2, 1/2)$
2	0.33036	0.073661
4	0.33743	0.073656
exact	0.3375	0.073671

Table 1.

REFERENCES

1. H.L. Langhaar and S.C. Chu, "Piecewise Polynomials and the Partition Method for Ordinary Differential Equations," Developments in Theoretical and Applied Mechanics, Vol 8, Pergamon Press, Oxford and New York, 1970.
2. S. Timoshenko and J. Goodier, Theory and Elasticity, 2nd ed., McGraw-Hill, New York, 1951.

ADAPTIVE, SELF-VALIDATING NUMERICAL QUADRATURE

George F. Corliss

*Department of Mathematics, Statistics, and Computer Science
Marquette University*

L. B. Rall

*Mathematics Research Center
University of Wisconsin-Madison*

Abstract. Integrals of a function of a single variable can be expressed as the sum of a numerical quadrature rule and a remainder term. The quadrature rule is a linear combination of function values and weights, or the integral of a Taylor polynomial, while the remainder term depends on some derivative of the integrand evaluated at an unknown point in the interval of integration. Numerical quadrature is made self-validating by using interval computation to capture both the roundoff and truncation errors made when using a given rule. Necessary derivatives can be generated automatically by using well-known recurrence relations for Taylor coefficients. In order for quadrature methods of this type to be accurate (in the sense that small intervals are produced) and efficient (to obtain results of given accuracy in a reasonably short time), an accurate scalar product and an adaptive strategy are required. The necessary scalar product and support for interval arithmetic are provided in Pascal-SC (for microcomputers) and ACRITH (for IBM 370 computers). The adaptive strategy chooses the subintervals of integration and the order of the quadrature formula in each subinterval on the basis of guaranteed, rather than estimated, information about the error of numerical integration in each subinterval. The program described in this report implements standard Newton-Cotes, Gaussian, and Taylor series methods for numerical integration. Ways to handle singularities are discussed, and comparisons are given with a standard numerical integration method.

1. Requirements for Automatic Integration Algorithms. In [2], de Boor formulates fundamental requirements for an automatic algorithm for numerical approximation of the integral

$$(1.1) \quad If = I_{[a,b]}f = \int_a^b f(x)dx$$

of a function of a single real variable. Such an algorithm requires (i) the limits of integration a, b , (ii) access to a procedure for the evaluation of $f(x)$ for x in the interval of integration, (iii) tolerances α, ρ on the desired absolute and relative error, respectively, and (iv) a limit M on the number of function evaluations allowed.

As output, the algorithm should produce an estimate I^* for the value of If which satisfies

$$(1.2) \quad |If - I^*| \leq \max\{\alpha, \rho |If|\}.$$

Sponsored in part by the U. S. Army under Contract No. DAAG29-80-C-0041.

Furthermore, the algorithm should be *efficient*, computing as few function values as possible. It should also be *reliable*, which will be taken here to mean that either the desired accuracy (1.2) is guaranteed, or a message to the contrary is returned to the user, possibly with additional information about the cause of failure. As pointed out by de Boor [2], algorithms which use only values of $f(x)$ at a finite number of points cannot meet the above requirements in general; nevertheless, accurate and efficient automatic integration algorithms can be formulated for wide classes of integrands [3], [23].

This paper presents automatic quadrature algorithms which attain the goals of reliability and efficiency by use of automatic differentiation and interval computation. They make use of information about the integrand on entire subintervals of integration, rather than at a discrete set of points. The results combine the self-validating algorithms of Gray and Rall [8], [9], [10], and the notion of adaptive quadrature [2], [3], [23]. Adaptation is carried out on the basis of guaranteed, rather than estimated, bounds for the error of the approximate integration over each subinterval. Furthermore, the given algorithm has the ability to detect and handle certain types of singularities in the integrand, and even to verify nonexistence of the integral in some cases.

In the terminology of Rice [23], the method described here has the following features:

Interval Processor Component:

Variable order rule with remainder using interval arithmetic to give guaranteed bounds.

Bound Estimator Component:

Direct analysis.

Special Behavior Components:

Polynomials.
Roundoff level.
Singularities in derivatives.
Jump discontinuities.
Removable singularities.
 x^α -type singularities.
All are strictly validated.

Interval Collection Management Component:

Ordered list.
None discarded.

The method of this paper does not belong to the large family of 10^6 or so algorithms considered by Rice because of the use of interval computation and automatic differentiation, which were not considered in [23]. Details of the actual implementation of the algorithms presented here in an environment which supports interval computation will be given in §7. The next few sections describe the underlying methodology.

2. Self-validating Evaluation of Quadrature Formulas. Self-validation of numerical

computations is one of the basic motivations of interval analysis [1], [15], [16]. The goal is to obtain an interval which contains the desired result, be it real or set-valued. In the case of the integration problem (1.1), a self-validating interval method produces an interval $J = [c, d]$ which is guaranteed to contain the value If of the integral. The width of this interval inclusion will depend on uncertainties in the values of the integrand and the limits of integration, the roundoff error in the actual computation, and the truncation error appropriate to the method used. All of these quantities can be estimated in a tedious way by the techniques of classical error analysis, an effort which is unnecessary in the computational environment described below. However, once an interval $[c, d]$ containing I is found by whatever method, one has the following approximations to I and corresponding error bounds [21]:

$$(2.1) \quad I^* = \frac{1}{2}(c + d), \quad |If - I^*| \leq \frac{1}{2}(d - c),$$

for absolute error, or

$$(2.2) \quad I^* = \frac{2cd}{c + d}, \quad \left| \frac{If - I^*}{If} \right| \leq \left| \frac{d - c}{c + d} \right|,$$

for relative error, with $cd > 0$ in this case. It follows that (1.2) will be satisfied if an interval $J = [c, d]$ can be obtained with width $w(J) = d - c$ small enough so that $w(J) \leq 2\alpha$ and $w(J) \leq \rho[c + d]$.

First, the problem of finding an interval inclusion J of If will be considered. The basic method for interval integration by use of standard formulas for numerical quadrature or Taylor series was first described by Moore [14]. To illustrate Moore's idea, consider a standard interpolatory integration formula of the form

$$(2.3) \quad \int_a^b f(x)dx = \sum_{i=1}^n w_i f(x_i) + c_n h \cdot \frac{f^{(p)}(\xi)h^p}{p!},$$

where $h = (b - a)/n$, and $a < \xi < b$. A formula of type (2.3) will be called a quadrature formula of order p on n points. The ordinary Gauss and Newton-Cotes integration formulas follow this pattern [7].

It should be noted that integration formulas such as (2.3) give the *exact* value of the integral If of functions which are differentiable p times. The only difficulty is that the value of ξ is unknown. For practical computation, it is thus customary to express (2.3) as the sum of a *rule*

$$(2.4) \quad r_n f = \sum_{i=1}^n w_i f(x_i)$$

of numerical integration, and a (truncation) *error term*

$$(2.5) \quad e_n f = c_n h \cdot f_p(\xi, h).$$

where

$$(2.6) \quad f_p(\xi, h) = \frac{f^{(p)}(\xi)h^p}{p!}$$

denotes the Taylor coefficient of order p in the expansion of $f(\xi + h)$. It is usual to compute $I^* = r_n f$ to approximate the value of the integral, and to estimate $e_n f$ somehow. Of course, if f is a polynomial of degree $p - 1$ or less, then $e_n f \equiv 0$, and $I f = r_n f$.

A self-validating computation of the rule $r_n f$ of numerical integration is straightforward in an environment which provides interval arithmetic and monotone interval inclusions of the library functions used in the evaluation of $f(x)$ for a given x . Let \mathbf{S} denote the screen of floating-point numbers available, and \mathbf{IS} the corresponding set of closed intervals $[u, v]$, $u, v \in \mathbf{S}$. If f is evaluated on an interval $X \in \mathbf{IS}$ using interval arithmetic and library functions, then the result is the *natural interval inclusion* $F(X)$ of f on X such that

$$(2.7) \quad f(X) = \{f(x) \mid x \in X\} \subseteq F(X).$$

[15], [16]. If W_i, X_i respectively denote the smallest intervals in \mathbf{IS} which contain the real numbers w_i, x_i , that is, $W_i = [\nabla w_i, \Delta w_i]$, $X_i = [\nabla x_i, \Delta x_i]$, where ∇, Δ denote the monotone downward and upward roundings from the real numbers \mathbf{R} to \mathbf{S} [11], then the inclusion

$$(2.8) \quad r_n f \in R_n f = \sum_{i=1}^n W_i F(X_i)$$

is guaranteed, and the computation of R_n can be done automatically.

An automatic, self-validating computation of the error term (2.5) requires an additional ingredient. This consists of subroutines for the generation of the Taylor coefficients $f_p(x, h)$ of f . These use well-known recurrence relations for the arithmetic operations and library functions used to evaluate $f(x)$ for given x [6], [15], [16], [19]. A suitable computational environment provides these routines. Corliss and Chang [5] have shown that the calculation of $f_0(x, h), f_1(x, h), \dots, f_p(x, h)$ requires about

$$(2.9) \quad t = ap^2 + bp + c$$

units of time, where a depends on the number of multiplications, divisions, and calls to library functions in the computation of f , and b depends on the number of additions and subtractions required. In any case, interval evaluation of exactly the same recurrence relations yields the corresponding interval inclusions $F_0(X, H), \dots, F_p(X, H)$ such that

$$(2.10) \quad f_k(x, h) \in F_k(X, H), \quad k = 0, 1, \dots, p,$$

for all $x \in X$ and $h \in H$ [15], [16]. Thus, the desired interval inclusion

$$(2.11) \quad e_n f \in E_n f = C_n H \cdot F_p(X, H)$$

can be computed automatically, given intervals $C_n, H, X \in \mathbf{IS}$ such that $c_n \in C_n, h \in H$, and $[a, b] \subset X$. (It is assumed that $a \leq b$; the contrary case can be handled easily.) It follows from the above that

$$(2.12) \quad If \in R_n f + E_n f = [c, d],$$

a formula for automatic, self-validating numerical quadrature. Once again, if f is a polynomial of degree $p - 1$ or less, then $F_p(X, H) \equiv [0, 0]$, so that $If \in R_n f$, and the width of $[c, d]$ depends only on the roundoff error in the calculation of $r_n f$.

This approach was the basis of an actual computer program [9], which met the accuracy criteria (1.2), if possible, by choosing H sufficiently small [18]. However, the problem of efficiency was not addressed.

Instead of splitting the numerical integration formula (2.3) into a rule of numerical integration (2.4) and an error term (2.5), it will be helpful later to consider it to represent the scalar product of the augmented *function-value vector*

$$(2.13) \quad \mathbf{f} = (f(x_1), \dots, f(x_n), f_p(\xi, h)),$$

and the augmented *weight vector*

$$(2.14) \quad \mathbf{w} = (w_1, \dots, w_n, c_n h),$$

which is independent of the integrand f and depends only on the specific formula (2.3) used. Thus,

$$(2.15) \quad I = \mathbf{w} \cdot \mathbf{f} \in \mathbf{W} \cdot \mathbf{F} = J,$$

where

$$(2.16) \quad \mathbf{F} = (F(X_1), \dots, F(X_n), F_p(X, H))$$

and

$$(2.17) \quad \mathbf{W} = (W_1, \dots, W_n, C_n H)$$

are the corresponding interval inclusions of \mathbf{f}, \mathbf{w} . This allows the computation to use recently developed methods for highly accurate calculation of real and interval scalar products [12]. This results in a considerable decrease in width due to roundoff error in the computed value of J .

The integration formula (2.3) can be interpreted as a *single-panel* rule, or as a *multi-panel* rule, meaning that a simpler formula on m points is applied k times to the corresponding number of subintervals of $X = [a, b]$, with $n \leq km$. Denoting the subintervals of X by $X_i, i = 1, 2, \dots, k$, this means that an integration formula

$$(2.18) \quad \int_{X_i} f(x) dx = \sum_{j=1}^m w_{ij} f(x_{ij}) + c_{im} h_i \cdot f_p(\xi_i, h_i),$$

holds in each subinterval, where $h_i = w(X_i)$. It has been shown [18] that

$$(2.19) \quad \sum_{i=1}^k c_{im} w(X_i) F^{(p)}(X_i) \cdot \frac{w(X_i)^p}{p!} \subseteq c_n w(X) F^{(p)}(X) \cdot \frac{w(X)^p}{p!}.$$

In addition to the decrease in width of $F_p(X, w(X))$ by a factor of $w(X)^p$ as $w(X)$ becomes small, the width of $F^{(p)}(X)$ will overestimate the width of $f^{(p)}(X)$ by less as $w(X) \rightarrow 0$ for $f^{(p)}(x)$ continuous [15]. Thus, the gain in calculating the error terms over smaller subintervals can be substantial. Roundoff error in adding a number of interval inclusions of one-panel rules (2.18) can again be reduced considerably by expressing the result as the scalar product of the extended augmented function-value vector

$$(2.20) \quad \mathbf{F} = (F(X_{11}, \dots, F(X_{1m}), F_p(X_1, H_1), \dots, F(X_{k1}), \dots, F(X_{km}), F_p(X_k, H_k))$$

with the extended augmented weight vector

$$(2.21) \quad \mathbf{W} = (w_{11}, \dots, w_{1m}, C_{1m}H_1, \dots, w_{k1}, \dots, w_{km}, C_{km}H_k).$$

3. Taylor Series Methods. The seminal paper by Moore [14] also provides the basis for self-validating numerical integration by the use of Taylor series, although the techniques presented by him in this case are directed toward the solution of the initial-value problem for ordinary differential equations. For numerical integration, Taylor series are more appropriate than fixed quadrature formulas for interval-valued endpoints of intervals of integration, as will be discussed in §6. Furthermore, Taylor series support the rigorous approach to automatic recognition and treatment of singularities described in §5.

Of course, one could consider (1.1) to be the solution $If = y(b)$ of the initial-value problem

$$(3.1) \quad y'(x) = f(x), \quad y(a) = 0,$$

and apply Moore's methods directly. However, since $f(x)$ in (3.1) is independent of y , unlike the usual case in differential equations, it is simpler to use the capability to generate a segment of the Taylor series and the interval remainder term automatically to perform a self-validating calculation of the desired integral.

In particular, instead of expanding the solution $y(x)$ of (3.1) at $x = a$ as in the case of a differential equation, it is advantageous to expand $f(x)$ at the midpoint $c = (a + b)/2$ of the interval $X = [a, b]$ of integration. It will be assumed that the integrand f has $p \geq 0$ derivatives in the interval of integration. For $h = (b - a)/2$, one has

$$(3.2) \quad \begin{aligned} f(x) = & f(c) + f'(c)(x - c) + f''(c) \frac{(x - c)^2}{2!} + \dots + f^{(n-1)}(c) \frac{(x - c)^{n-1}}{(n-1)!} \\ & + f^{(n)}(\xi) \frac{(x - c)^n}{n!}, \end{aligned}$$

can be computed automatically, given intervals $C_n, H, X \in \mathbf{IS}$ such that $c_n \in C_n, h \in H$, and $[a, b] \subset X$. (It is assumed that $a < b$; the contrary case can be handled easily.) It follows from the above that

$$(2.12) \quad If \in R_n f \cup E_n f \subset [c, d],$$

a formula for automatic, self-validating numerical quadrature. Once again, if f is a polynomial of degree $p - 1$ or less, then $F_p(X, H) = 0, 0$, so that $If \in R_n f$, and the width of $[c, d]$ depends only on the roundoff error in the calculation of $r_n f$.

This approach was the basis of an actual computer program [9], which met the accuracy criteria (1.2), if possible, by choosing H sufficiently small [18]. However, the problem of efficiency was not addressed.

Instead of splitting the numerical integration formula (2.3) into a rule of numerical integration (2.4) and an error term (2.5), it will be helpful later to consider it to represent the scalar product of the augmented *function-value vector*

$$(2.13) \quad \mathbf{f} = (f(x_1), \dots, f(x_n), f_p(\xi, h)),$$

and the augmented *weight vector*

$$(2.14) \quad \mathbf{w} = (w_1, \dots, w_n, c_n h),$$

which is independent of the integrand f and depends only on the specific formula (2.3) used. Thus,

$$(2.15) \quad I \in \mathbf{w} \cdot \mathbf{f} \in \mathbf{W} \cdot \mathbf{F} \in J,$$

where

$$(2.16) \quad \mathbf{F} = (F(X_1), \dots, F(X_n), F_p(X, H))$$

and

$$(2.17) \quad \mathbf{W} = (W_1, \dots, W_n, C_n H)$$

are the corresponding interval inclusions of \mathbf{f}, \mathbf{w} . This allows the computation to use recently developed methods for highly accurate calculation of real and interval scalar products [12]. This results in a considerable decrease in width due to roundoff error in the computed value of J .

The integration formula (2.3) can be interpreted as a *single-panel* rule, or as a *multi-panel* rule, meaning that a simpler formula on m points is applied k times to the corresponding number of subintervals of $X = [a, b]$, with $n = km$. Denoting the subintervals of X by $X_i, i = 1, 2, \dots, k$, this means that an integration formula

$$(2.18) \quad \int_{X_i} f(x) dx = \sum_{j=1}^m w_{ij} f(x_{ij}) + c_{im} h_i \cdot f_p(\xi_i, h_i),$$

holds in each subinterval, where $h_i = w(X_i)$. It has been shown [18] that

$$(2.19) \quad \sum_{i=1}^k c_{im} w(X_i) F^{(p)}(X_i) \cdot \frac{w(X_i)^p}{p!} \subseteq c_n w(X) F^{(p)}(X) \cdot \frac{w(X)^p}{p!}.$$

In addition to the decrease in width of $F_p(X, w(X))$ by a factor of $w(X)^p$ as $w(X)$ becomes small, the width of $F^{(p)}(X)$ will overestimate the width of $f^{(p)}(X)$ by less as $w(X) \rightarrow 0$ for $f^{(p)}(x)$ continuous [15]. Thus, the gain in calculating the error terms over smaller subintervals can be substantial. Roundoff error in adding a number of interval inclusions of one-panel rules (2.18) can again be reduced considerably by expressing the result as the scalar product of the extended augmented function-value vector

$$(2.20) \quad \mathbf{F} = (F(X_{11}, \dots, F(X_{1m}), F_p(X_1, H_1), \dots, F(X_{k1}), \dots, F(X_{km}), F_p(X_k, H_k))$$

with the extended augmented weight vector

$$(2.21) \quad \mathbf{W} = (w_{11}, \dots, w_{1m}, C_{1m}H_1, \dots, w_{k1}, \dots, w_{km}, C_{km}H_k).$$

3. Taylor Series Methods. The seminal paper by Moore [14] also provides the basis for self-validating numerical integration by the use of Taylor series, although the techniques presented by him in this case are directed toward the solution of the initial-value problem for ordinary differential equations. For numerical integration, Taylor series are more appropriate than fixed quadrature formulas for interval-valued endpoints of intervals of integration, as will be discussed in §6. Furthermore, Taylor series support the rigorous approach to automatic recognition and treatment of singularities described in §5.

Of course, one could consider (1.1) to be the solution $If = y(b)$ of the initial-value problem

$$(3.1) \quad y'(x) = f(x), \quad y(a) = 0,$$

and apply Moore's methods directly. However, since $f(x)$ in (3.1) is independent of y , unlike the usual case in differential equations, it is simpler to use the capability to generate a segment of the Taylor series and the interval remainder term automatically to perform a self-validating calculation of the desired integral.

In particular, instead of expanding the solution $y(x)$ of (3.1) at $x = a$ as in the case of a differential equation, it is advantageous to expand $f(x)$ at the midpoint $c = (a + b)/2$ of the interval $X = [a, b]$ of integration. It will be assumed that the integrand f has $p \geq 0$ derivatives in the interval of integration. For $h = (b - a)/2$, one has

$$(3.2) \quad \begin{aligned} f(x) = & f(c) + f'(c)(x - c) + f''(c) \frac{(x - c)^2}{2!} + \dots + f^{(n-1)}(c) \frac{(x - c)^{n-1}}{(n-1)!} \\ & + f^{(n)}(\xi) \frac{(x - c)^n}{n!}. \end{aligned}$$

for $n \leq p$, where $\xi \in X$ is between c and x . Let $F^{(n)}$ be an interval inclusion of $f^{(n)}$ on X . Then, $f^{(n)}(\xi) \in F^{(n)}(X)$, which is an interval-valued constant. From (3.2),

$$(3.3) \quad f(x) \in f(c) + f'(c)(x-c) + \dots + f^{(n-1)}(c) \frac{(x-c)^{n-1}}{(n-1)!} + F^{(n)}(X) \frac{(x-c)^n}{n!}.$$

Let

$$(3.4) \quad g(x) = f(c)(x-c) + f'(c) \frac{(x-c)^2}{2!} + \dots + f^{(n-1)}(c) \frac{(x-c)^n}{n!}$$

be an indefinite integral of the Taylor polynomial of degree $n-1$ of $f(x)$. Then,

$$(3.5) \quad \int_a^b f(x) dx \in g(x) \Big|_a^b + F^{(n)}(X) \frac{(x-c)^{n+1}}{(n+1)!} \Big|_a^b \subseteq J_n,$$

where

$$(3.6) \quad J_n = 2 \sum_{\substack{i=0 \\ i \text{ even}}}^{n-1} F^{(i)}(C) \frac{H^{i+1}}{(i+1)!} + \begin{cases} \left[F^{(n)}(X) \frac{H^{n+1}}{(n+1)!} - F^{(n)}(X) \frac{H^{n+1}}{(n+1)!} \right], & \text{for } n \text{ odd,} \\ 2F^{(n)}(X) \frac{H^{n+1}}{(n+1)!}, & \text{for } n \text{ even.} \end{cases}$$

Note that subtraction does not "cancel" equal intervals in general. One has $[u, v] - [u, v] = [u-v, v-u] \neq [0, 0]$ unless $u = v$, in which case the interval consists of a single point [1], [15], [16]. If the series were expanded at $x = a$ instead of $x = c$, then the width of the interval remainder terms would be increased by a factor of 2^{n+1} .

Formulas (3.5)-(3.6) resemble (2.12) for ordinary quadrature rules, with the evaluation of the integrand at n points replaced by its value and the values of its first $n-1$ derivatives at a single point. If f is a polynomial of degree $n-1$ or less, then $F^{(n)}(X) \equiv [0, 0]$, and only roundoff error effects the width of J_n .

J_n can be computed directly from the automatically generated interval Taylor coefficients $F_k(C, H)$ and $F_k(X, H)$ of f , $k = 0, 1, \dots, n \leq p$.

Since

$$(3.7) \quad If = \int_a^b f(x) dx \in J_n, \quad n = 0, 1, \dots, p,$$

the *intersection principle* [8], [9] can be used. One has

$$(3.8) \quad If \in \bigcap_{n=0}^p J_n.$$

This intersection can be calculated as the corresponding interval Taylor coefficients of the integrand are generated:

$$(3.9) \quad \begin{cases} I_0 = J_0, & \text{(a Riemann sum).} \\ I_n = I_{n-1} \cap J_n, & n = 1, 2, \dots, p. \end{cases}$$

This provides a means to determine the highest useful term of the Taylor expansion, since the calculation can be terminated when effective decrease in the widths of the intervals $\{I_n\}$ ceases, or when the desired tolerance is met.

As in the case of (2.12), formula (3.6) can be evaluated as the scalar product of the vectors

$$(3.10) \quad \mathbf{F} = (F_0(C, H), F_1(C, H), \dots, F_{n-1}(C, H), T_n(C, H)),$$

where $T_n(C, H) = F_n(C, H) - F_n(C, H)$ or $T_n(C, H) = 2F_n(C, H)$ according as n is odd or even, and the vector

$$(3.11) \quad \mathbf{W} = (W_0, W_1, \dots, W_{n-1}, W_n),$$

where

$$(3.12) \quad W_k = \frac{H}{k+1}, \quad k = 0, 1, \dots, n.$$

The width of J_n can also be reduced somewhat if the expansion is about $C = [c, c]$, $c \in S$. In this case, the interval stepsize H might have to be increased a very small amount to maintain inclusion of the subinterval of integration. In §5, it will be noted that expansion about an endpoint, as in (3.1), or some other point in the interval of integration, can be helpful if the integrand has removable singularities.

4. Adaptive Strategies. The computer program INTE [9] demonstrated the reality of automatic, self-validating numerical quadrature using Newton-Cotes and Gaussian integration formulas in the way discussed in §2. (Euler-Maclaurin integration was added to the capabilities of INTE later [10].) However, there was no attempt to address the problem of efficiency, a defect remedied in the program described later. In order to satisfy the accuracy criterion (1.2), if possible, INTE simply divided the interval of integration into a sufficient number of equal subintervals [9], [18]. By contrast, popular numerical integration packages such as CADRE [3] and QUADPACK [17] use information obtained about the behavior of the integrand to attempt to reduce the number of function evaluations to a minimum. The method given here uses similar strategies, except that estimates of the error based on evaluation of the integrand at a finite set of points are replaced by guaranteed bounds. This eliminates the need for "safety factors" [23].

Adaptive strategies fall into the categories of *order* adaptation, which relates to the choice of the formula used in each subinterval, and *subinterval* adaptation, which determines how the original interval of integration is broken up into subintervals. Order

for $n = p$, where $\xi \in X$ is between c and x . Let $F^{(n)}$ be an interval inclusion of $f^{(n)}$ on X . Then, $f^{(n)}(\xi) \in F^{(n)}(X)$, which is an interval-valued constant. From (3.2),

$$(3.3) \quad f(x) = f(c) + f'(c)(x-c) + \dots + f^{(n-1)}(c) \frac{(x-c)^{n-1}}{(n-1)!} + F^{(n)}(X) \frac{(x-c)^n}{n!}.$$

Let

$$(3.4) \quad g(x) = f(c)(x-c) + f'(c) \frac{(x-c)^2}{2!} + \dots + f^{(n-1)}(c) \frac{(x-c)^n}{n!}$$

be an indefinite integral of the Taylor polynomial of degree $n-1$ of $f(x)$. Then,

$$(3.5) \quad \int_a^b f(x) dx = g(x) \Big|_a^b + F^{(n)}(X) \frac{(x-c)^{n+1}}{(n+1)!} \Big|_a^b = J_n,$$

where

$$(3.6) \quad J_n = 2 \sum_{\substack{i=0 \\ i \text{ even}}}^{n-1} F^{(i)}(C) \frac{H^{i+1}}{(i+1)!} + \begin{cases} \left[F^{(n)}(X) \frac{H^{n+1}}{(n+1)!} - F^{(n)}(X) \frac{H^{n+1}}{(n+1)!} \right], & \text{for } n \text{ odd,} \\ 2F^{(n)}(X) \frac{H^{n+1}}{(n+1)!}, & \text{for } n \text{ even.} \end{cases}$$

Note that subtraction does not "cancel" equal intervals in general. One has $[u, v] - [u, v] = [u-v, v-u] \neq [0, 0]$ unless $u = v$, in which case the interval consists of a single point [1], [15], [16]. If the series were expanded at $x = a$ instead of $x = c$, then the width of the interval remainder terms would be increased by a factor of 2^{n+1} .

Formulas (3.5)-(3.6) resemble (2.12) for ordinary quadrature rules, with the evaluation of the integrand at n points replaced by its value and the values of its first $n-1$ derivatives at a single point. If f is a polynomial of degree $n-1$ or less, then $F^{(n)}(X) = [0, 0]$, and only roundoff error effects the width of J_n .

J_n can be computed directly from the automatically generated interval Taylor coefficients $F_k(C, H)$ and $F_k(X, H)$ of f , $k = 0, 1, \dots, n = p$.

Since

$$(3.7) \quad If = \int_a^b f(x) dx = J_n, \quad n = 0, 1, \dots, p,$$

the *intersection principle* [8], [9] can be used. One has

$$(3.8) \quad If = \bigcap_{n=0}^p J_n.$$

This intersection can be calculated as the corresponding interval Taylor coefficients of the integrand are generated:

$$(3.9) \quad \begin{cases} I_0 = J_0, & \text{(a Riemann sum),} \\ I_n = I_{n-1} \cap J_n, & n = 1, 2, \dots, p. \end{cases}$$

This provides a means to determine the highest useful term of the Taylor expansion, since the calculation can be terminated when effective decrease in the widths of the intervals $\{I_n\}$ ceases, or when the desired tolerance is met.

As in the case of (2.12), formula (3.6) can be evaluated as the scalar product of the vectors

$$(3.10) \quad \mathbf{F} = (F_0(C, H), F_1(C, H), \dots, F_{n-1}(C, H), T_n(C, H)),$$

where $T_n(C, H) = F_n(C, H) - F_n(C, H)$ or $T_n(C, H) = 2F_n(C, H)$ according as n is odd or even, and the vector

$$(3.11) \quad \mathbf{W} = (W_0, W_1, \dots, W_{n-1}, W_n),$$

where

$$(3.12) \quad W_k = \frac{H}{k+1}, \quad k = 0, 1, \dots, n.$$

The width of J_n can also be reduced somewhat if the expansion is about $C = [c, c]$, $c \in S$. In this case, the interval stepsize H might have to be increased a very small amount to maintain inclusion of the subinterval of integration. In §5, it will be noted that expansion about an endpoint, as in (3.1), or some other point in the interval of integration, can be helpful if the integrand has removable singularities.

4. Adaptive Strategies. The computer program INTE [9] demonstrated the reality of automatic, self-validating numerical quadrature using Newton-Cotes and Gaussian integration formulas in the way discussed in §2. (Euler-Maclaurin integration was added to the capabilities of INTE later [10].) However, there was no attempt to address the problem of efficiency, a defect remedied in the program described later. In order to satisfy the accuracy criterion (1.2), if possible, INTE simply divided the interval of integration into a sufficient number of equal subintervals [9], [18]. By contrast, popular numerical integration packages such as CADRE [3] and QUADPACK [17] use information obtained about the behavior of the integrand to attempt to reduce the number of function evaluations to a minimum. The method given here uses similar strategies, except that estimates of the error based on evaluation of the integrand at a finite set of points are replaced by guaranteed bounds. This eliminates the need for "safety factors" [23].

Adaptive strategies fall into the categories of *order* adaptation, which relates to the choice of the formula used in each subinterval, and *subinterval* adaptation, which determines how the original interval of integration is broken up into subintervals. Order

adaptation is somewhat simpler than subinterval adaptation, and will be considered first. For methods based either on standard quadrature formulas or Taylor series, *order zero* refers to the interval Riemann sum $F(X) \cdot w(X)$, which always contains $\int_X f(x)dx$ [4].

Given a suite of numerical integration formulas of the form (2.3), suppose that $X \in \mathbf{IS}$ is the current subinterval of integration. Specifically, suppose that the given rules are of order i for $i = 0, 1, \dots, k$ on n_i points. Once the interval Taylor coefficients $F_i(X, H)$ have been formed, then the order $p \leq k$ of the most accurate rule can be chosen to be the value of i for which the width of the error term

$$(4.1) \quad E_{n,i}f = C_{n,i}H \cdot F_i(X, H)$$

is minimum, $i = 0, 1, \dots, k$.

Alternatively, the actual approximate integrals J_i could be examined, but this requires more computation than use of the error terms alone. Suppose that μ denotes the maximum value of the width $w(F(C))$ of the interval evaluation of f at a node C of the integration formula being used. The total cost can be taken to be proportional to $n_i \cdot w(J_i)$, where n_i is the number of function evaluations. The width $w(J_i)$ can be estimated by $\mu + w(E_{n,i}f)$, and thus i can be chosen as the minimizer of

$$(4.2) \quad m(i) = \min\{w(J_0), n_i(\mu + w(E_{n,i}f))\}.$$

Thus, more nodes are used in a given subinterval only if a significant reduction in width of the approximate integral results. It should also be noted that a given integration rule can be used in several integration formulas having remainder terms of different orders. For example, Stroud and Secrest [24] give error terms for Gaussian integration rules on n nodes which have orders $1 \leq p \leq 2n$. In certain cases, the error terms corresponding to smaller than maximum p can be narrower (for example, for highly oscillatory integrands), and the use of these formulas instead of the standard ones will minimize $w(J_n)$.

In the case of Taylor series, the intersection (3.9) procedure can result in approximations I_n which are considerably better than J_n given by (3.6), which can be viewed as the sum of a rule and an error term. The intervals I_n are monotone decreasing in width, so the increase in accuracy has to be balanced against the cost in time (2.9) of generating more terms of the series. The constants a, b, c in (2.9) can always be determined for a given integrand, so the corresponding heuristic can be based on the function

$$(4.3) \quad \theta(n) = w(I_n) \cdot (an^2 + bn + c).$$

Generation of the Taylor series can be stopped when

- (i) $I_{n-2} = I_{n-1} = I_n$,
- (ii) $\theta(n-2) \leq \theta(n)$, or
- (iii) $n = p$.

Because of the difference between the remainder terms in (3.6) for odd and even n , it is prudent to calculate at least one extra value of J_n . However, the conditions above guarantee that no more than two series terms will be computed beyond the one which gives the narrowest I_n .

Another important strategy for order adaptation involves the case that the integrand has singularities in a certain derivative in the given subinterval. This will be discussed more fully in the next section. If a certain derivative cannot be evaluated, this will be detected, and the method will be restricted to rules or orders of Taylor expansion which use only the derivatives which can be evaluated.

The strategy for subinterval adaptation retains all subintervals. At each step, the subinterval which makes the largest contribution to the width of J is processed by breaking it into further subintervals. This processing continues until

- (i) $w(J)$ is small enough to satisfy the accuracy requirement (1.2),
- (ii) the noise inherent in function evaluation limits further reduction of $w(J)$, or
- (iii) more than the maximum number M of function evaluations have been performed.

The second termination criterion in this list is particularly important. If the noise in the function evaluation is large relative to the accuracy requested, eventually the width of the truncation error Ef is made so small that it adds nothing to the width of the rule Rf alone. At this point, further increase in accuracy is not possible. Without the guaranteed bounds provided by the interval computation, many standard methods cannot recognize when this point has been reached, and the calculation should be terminated. Malcolm and Simpson [13] observe that the strategy of processing the worst subinterval results in local errors of roughly equal magnitude. Rice [23] calls this an ordered list interval collection management component, and lists some advantages and disadvantages which will be discussed in more detail in §7.

5. Treatment of Singularities. Among the quadrature routines which provide estimates for If , the more successful ones have special provisions to handle and perhaps recognize certain types of singularities. QUADPACK [17], for example, can handle integrals of the form

$$(5.1) \quad \int_a^b (x-a)^\alpha (b-x)^\beta v(x) f(x) dx,$$

where $\alpha, \beta > -1$ and $v(x) = 1, \log(x-a), \log(b-x)$, or $\log(x-a) \log(b-x)$, *provided* the *user* supplies α, β , and the form of v . CADRE [3] attempts to detect and verify the presence of jump discontinuities or x^α -type singularities in the integrand. The program described here attempts to recognize and handle automatically

- (1) singularities in derivatives of f ,
- (2) some removable singularities,
- (3) jump discontinuities, and
- (4) some algebraic singularities at endpoints.

Since guaranteed bounds are computed for If , the task of such a program is more difficult than for methods which yield only an estimate. For example. Gaussian quadrature can be used in the neighborhood of an endpoint singularity, because the integrand need not be evaluated at the endpoint. To give guaranteed bounds, however, requires that the function, or some of its derivatives, be evaluated on the entire interval. Hence, the price

of the guarantee is a restriction on the applicability of the program. Either it verifies that the integrand is in its domain of applicability, or, if it cannot deliver guaranteed bounds, it notifies the user with an indication of the difficulty.

An example will indicate what can go wrong at endpoints. For example,

$$(5.2) \quad I_1 f = \int_0^\pi \sqrt{\pi - x} \, dx = \frac{2}{3} \pi^{\frac{3}{2}}.$$

could be presented to the computer as

$$(5.3) \quad I_2 f = \int_0^{[\nabla \pi, \Delta \pi]} \sqrt{[\nabla \pi, \Delta \pi] - x} \, dx.$$

Although the original problem (5.2) is well-behaved, the interval integral in (5.3) contains integrals such as

$$(5.4) \quad \int_0^{\Delta \pi} \sqrt{\nabla \pi - x} \, dx,$$

and consequently does not exist. Hence, the correct response is that (5.3) cannot be evaluated on the entire interval of integration, just on $[0, \nabla \pi]$, for example. This suggests that standard numerical quadrature routines, which avoid evaluation of the integrand at endpoints, can be fooled into returning values for integrals such as

$$(5.5) \quad \int_0^\pi \sqrt{\pi - \epsilon - x} \, dx$$

for ϵ sufficiently small, instead of informing the user of possible difficulty. The types of singularities which can be handled by the techniques presented here will now be described.

5.1. Singularities in derivatives of f .

The integrand f may not have enough derivatives for some rules which the integration program can apply. In the process of automatic generation of interval Taylor coefficients on an interval X , the nonexistence of a derivatives of f beyond a certain order is detected and reported. As long as f itself can be evaluated on the entire interval of integration, the order adaptation strategy handles singularities in the derivatives of f . For example, consider the problem

$$(5.6) \quad I f = \int_0^4 \sqrt{x} \, dx = \frac{16}{3}.$$

The first derivative f' is undefined at $x = 0$, so the only rule that can be applied to the entire interval of integration is the Riemann sum

$$(5.7) \quad \int_0^4 \sqrt{x} \, dx \in \left[\min_{[0,4]} \sqrt{x}, \max_{[0,4]} \sqrt{x} \right] \cdot 4 = [0, 8].$$

At this point the subinterval adaptive strategy takes over. At each step, the subinterval containing 0 requires a Riemann sum, and thus is frequently selected for further processing because of its width. All other subintervals can be processed using higher-order rules. This strategy applies to singularities in f' which occur anywhere in the interval of integration, and not just at endpoints. It also works for singularities in higher derivatives, whose presence might not even be known to the user. Once a singularity of this type has been confined to a sufficiently small subinterval, it is possible to meet reasonable requirements for accuracy, with self-validation.

5.2. Jump discontinuities

If the integrand is given by an ordinary mathematical expression, then it is difficult to represent a function which has jump discontinuities, and yet can be evaluated at every point in the interval of integration. However, the user can supply a subroutine for evaluation of the integrand which produces jump discontinuities. These appear to the program as singularities in f' . On any subinterval which contains a jump, only the Riemann sum is available, and its width will lead to frequent selection of this subinterval for further processing, as before. No special algorithms are needed in this case.

This behavior is similar to CADRE [b2]. Upon recognizing a jump, CADRE subdivides the interval and uses a low-order rule.

5.3. Some removable singularities

The Taylor series method permits handling of some removable singularities. Consider the problem

$$(5.8) \quad If = \int_0^\pi \frac{\sin x}{x} dx,$$

in which the integrand has a removable singularity at $x = 0$. In this case, the integrand cannot be evaluated directly on any interval containing 0, but f can be expanded at $x = 0$ using l'Hôpital's rule, which can be applied automatically [6]. A short Taylor series with remainder for f at 0 is sufficient to bound If near 0, and the rest of the interval of integration is processed in the normal manner.

5.4. Some algebraic singularities at endpoints

Suppose that the integrand has the form

$$(5.9) \quad f(x) = (a - x)^{-s} \phi(x),$$

where $\phi(x)$ is analytic at $x = a$. If c is chosen closer to a than any other singularity, then the series for f expanded at $x = a$ is asymptotic to the series for $v(x) = (a - x)^{-s}$. The Taylor coefficients $v_i = v_i(a, h)$ of v satisfy the recurrence relation

$$(5.10) \quad v_{i+1} = v_i \left(1 + \frac{s-1}{i} \right) \frac{h}{R_c},$$

$$s = \left(\frac{v_{i+1}}{v_i} \frac{h}{R_c} - 1 \right) i + 1.$$

where R_c is the radius of convergence of the series. If f cannot be evaluated on $[a, b]$, then one attempts to find constants K_L, K_R, s_L, s_R such that

$$(5.11) \quad K_L(a-x)^{-s_L} \leq f(x) \leq K_R(a-x)^{-s_R},$$

for x near a . If such constants can be found and (5.11) validated, then we proceed. If $s_L > 1$, then the program can guarantee that If does not exist. We know of no other numerical quadrature routine which can *validate* nonexistence of an integral. If $s_R < 1$, then

$$(5.12) \quad \frac{K_L}{1-s_L}(a-a')^{1-s_L} \leq \int_a^{a'} f(x)dx \leq \frac{K_R}{1-s_R}(a-a')^{1-s_R}.$$

These bounds can be made as tight as desired by taking a' close enough to a . The interval $[a, a']$ is placed on the list of subintervals, and processing continues on the subinterval $[a', b]$.

A singularity at b can be handled similarly.

If K_L, K_R, s_L, s_R cannot be found, or if

$$(5.13) \quad s_L < 1 \leq s_R,$$

then the integrand cannot be handled by this method, and a message to that effect is sent to the user.

These methods for recognizing and handling certain singularities extend the domain of applicability of the program to include many integrands which arise in applications. However, their usefulness is somewhat limited. For one thing, the location of the singularity must be known in advance, or guessed. For popular sets of test problems, guessing one or both endpoints usually works. For real problems, locations of singularities may be unknown. However, the method given here validates correct guesses. The endpoints to be investigated for the presence of singularities must be machine numbers, not intervals. If an interval-valued endpoint is even one machine number wide, then the problem contains integrands which are unbounded on a set of positive measure. If a singularity is in the interior of the interval of integration, then we can determine its location to within one or two machine numbers. From this, its contribution to If could be estimated, but the possibility that the integrand is unbounded on a set of positive measure cannot be eliminated. Since a validated answer cannot be produced in this case, an error return is selected. If the user knows that the singularity occurs at a machine number, then the integral should be calculated as the sum of two integrals, with the singularity at one endpoint of each.

6. Extension to Interval Values. The definition of the integral (1.1) has been extended to arbitrary interval-valued integrands f [4], [20]. For smooth, real-valued functions such as those considered in §2, the interval integral and the Riemann integral coincide. The concept of interval integration is useful in connection with integrands which have jump discontinuities or singularities of various kinds. The definition given in [4] can be extended to the case that the limits of integration are interval-valued:

$$(6.1) \quad \int_A^B f(x)dx = \left\{ \int_a^b f(x)dx \mid a \in A, b \in B \right\}.$$

for $A, B \in \mathbf{IR}$, the set of all finite intervals with real endpoints.

There are several reasons to want to be able handle interval endpoints of integration. First of all, some real numbers, such as π , cannot be represented exactly in \mathbf{S} , and have to be replaced by the corresponding small intervals such as $[\nabla\pi, \Delta\pi]$ in \mathbf{IS} . Secondly, the limits of integration may come from measurements other estimates, and thus are known to lie between certain limits even though their exact values are uncertain. Finally, one can be interested in bounds for the value of the integral (1.1) over ranges of values of limits of integration. In this case, rather than satisfy an accuracy criterion such as (1.2), it is usually desired to find an interval J containing (6.1) which is as small as possible.

Similar considerations apply to interval-valued integrands. The case which usually arises in practice is that the integrand f is a function not only of the independent variable x , but also several parameters c_1, c_2, \dots, c_m . For example, f could be a polynomial of degree $m - 1$ with coefficients determined by observations,

$$(6.2) \quad f(x) = C_1 + C_2x + \dots + C_mx^{m-1}, \quad C_i \in \mathbf{IS}, \quad i = 1, 2, \dots, m.$$

In general, given intervals C_1, C_2, \dots, C_m , it is natural to define

$$(6.3) \quad f(x; C_1, \dots, C_m) = \{f(x; c_1, \dots, c_m) \mid c_1 \in C_1, \dots, c_m \in C_m\}.$$

The natural interval inclusions $F_k(X, C_1, \dots, C_m)$ of f and its Taylor coefficients on an interval $X \in \mathbf{IS}$ are again obtainable on a computer by using interval computation and automatic differentiation. In particular, for an interval polynomial (6.2), $F_p(X, H) \equiv [0, 0]$ for $p \geq m$, just as in the real case. The definition

$$(6.4) \quad \int_A^B f(x; C_1, \dots, C_m) dx = \left\{ \int_A^B f(x, c_1, \dots, c_m) dx \mid c_1 \in C_1, \dots, c_m \in C_m \right\}$$

describes the type of integrals to which the methods of this paper apply. It is assumed that A, B and the coefficient intervals C_i all belong to \mathbf{IS} , and hence are machine-representable. If necessary, outward rounding can be used to obtain them from real intervals, preserving inclusion of the desired integral.

On the basis of (2.3), it is to be expected that the best results will be obtained for integrands f which are very smooth as functions of x . However, one can always compute the *interval Riemann sum* $F(X) \cdot w(X)$. This is self-validating, because

$$(6.5) \quad \int_X f(x) dx \subseteq F(X) \cdot w(X),$$

but is inaccurate and slow to converge [4], [20]. It is important to be able to confine bad behavior of the integrand to very small intervals for (6.5) to be useful.

The meaning of tolerance for problems with interval-valued endpoints requires some clarification. Consider the problem

$$(6.6) \quad \int_0^{[0,1]} \frac{dx}{1+x^2} = \left[\int_0^0 \frac{dx}{1+x^2}, \int_0^1 \frac{dx}{1+x^2} \right] = \left[0, \frac{\pi}{4} \right].$$

If we compute $J = [-0.004, 0.79]$, for example, then $w(J) = 0.794$, although the estimate for each endpoint is in error by less than 0.005. Hence, a requested tolerance must be large enough to accomodate the uncertainties which are inherent in the problem being solved.

The cases that one or both endpoints of integration are nondegenerate intervals in **IS** will now be considered. The possible situations will be denoted by IR, RI, and II, respectively, according as the lower, upper, or both endpoints of the interval of integration are interval-valued. One strategy for handling interval-valued endpoints is to allow the uncertainties in the locations of the endpoints to be carried over into uncertainties in the locations of the nodes, as in INTE [9]. For example, using a one-panel Simpson's rule on 4-decimal digit machine gives

$$(6.7) \quad \int_{[0,0.1]}^{[3.1,3.2]} f(x)dx \in \frac{(B-A)}{6} (F([0,0.1]) + 4F([1.55,1.65]) + F([3.1,3.2])) \\ - \frac{(B-A)^5}{2880} F^{iv}([0,3.2]),$$

where $A = [0,0.1]$, $B = [3.1,3.2]$. For $f(x) = \sin x$, (6.7) yields $[1.880, 2.214]$ using the accurate scalar product, while the correct answer is $[1.994, 2.000]$.

A better strategy is to concentrate the uncertainty at the ends:

$$(6.8) \quad \int_{[0,0.1]}^{[3.1,3.2]} f(x)dx = \int_{[0,0.1]}^{0.1} f(x)dx + \int_{0.1}^{3.1} f(x)dx + \int_{3.1}^{[3.1,3.2]} f(x)dx.$$

The middle part is handled as an RR integral as described in §§3-4, while the other two integrals, of the types IR and RI, respectively, will be treated in the manner to be described below. In this example, we can get $[1.984, 2.073]$. As the widths of the endpoints increase, the advantage of the second strategy becomes more pronounced.

The general case (6.3) can be expressed in terms of integrals of the above types. This case in turn splits into six subcases, according as A and B are (i) disjoint, (ii) overlapping, or (iii) one is contained in the other. More precisely, for $A = [A_L, A_R]$, $B = [B_L, B_R]$, suppose that $A_R \leq B_R$. Then we have:

Case 1. $AR \leq BL$ (disjoint interval endpoints). Here, the integral (6.3) can be written

$$(6.9) \quad \int_A^B f(x)dx = \int_{[A_L, A_R]}^{A_R} f(x)dx + \int_{A_R}^{B_L} f(x)dx + \int_{B_L}^{[B_L, B_R]} f(x)dx,$$

and thus is the sum of integrals of types IR, RR, and RI, respectively.

Case 2. $AL \leq BL \leq AR \leq BR$ (overlapping interval endpoints). Here,

$$(6.10) \quad \int_A^B f(x)dx = \int_{[A_L, B_L]}^{B_L} f(x)dx + \int_{[B_L, A_R]}^{[B_L, A_R]} f(x)dx + \int_{A_R}^{[A_R, B_R]} f(x)dx,$$

the sum of integrals of types IR, II, and RI.

Case 3. $BL < AL \leq AR \leq BR$ (A is properly contained in B). Here.

$$(6.11) \quad \int_A^B f(x)dx = - \int_{[B_L, A_L]}^{A_L} f(x)dx - \int_{[A_L, A_R]}^{[A_L, A_R]} f(x)dx + \int_{A_R}^{[A_R, B_R]} f(x)dx,$$

again the sum of integrals of types IR, II, and RI.

The other three cases are obtained for $B_R < A_R$ by reversing the rôles of A and B in the preceding.

Case 1 is undoubtedly the one which is encountered most often in practice. To illustrate Case 3, consider

$$(6.12) \quad \int_{[2,3]}^{[0,5]} f(x)dx = - \int_{[0,2]}^2 f(x)dx + \int_{[2,3]}^{[2,3]} f(x)dx + \int_3^{[3,5]} f(x)dx.$$

We are not aware of physical problems which give rise to integration problems other than Case 1 (except to bound the values of integrals over ranges of endpoints), but provision for these cases adds little to the machinery which is required to handle Case 1.

Integrals of types RI, IR, and II can be handled by Gauss or Newton-Cotes formulas with interval-valued nodes, but this leads to wide interval bounds. The method of Taylor series, as outlined in §4, applies as easily to the cases of one or both endpoints interval-valued as to RR type integration. Hence, the program uses Taylor polynomials for integrals of types RI, IR, and II, even if the user has chosen Gauss or Newton-Cotes formulas for use on type RR subintervals.

Using the same notation as in §3, $g(x)$ denotes an indefinite integral (3.4) of $f(x)$. The one-panel form of the various integration formulas are then:

Type RI:

$$(6.13) \quad \int_a^{[a,b]} f(x)dx \subseteq g(x) \Big|_a^{[a,b]} + F^{(n)}([a,b]) \frac{(x-c)^{n+1}}{(n+1)!} \Big|_a^{[a,b]},$$

Type IR:

$$(6.14) \quad \int_{[a,b]}^b f(x)dx \subseteq g(x) \Big|_{[a,b]}^b + F^{(n)}([a,b]) \frac{(x-c)^{n+1}}{(n+1)!} \Big|_{[a,b]}^b,$$

Type II:

$$(6.15) \quad \int_{[a,b]}^{[a,b]} f(x)dx \subseteq g(x) \Big|_{[a,b]}^{[a,b]} + F^{(n)}([a,b]) \frac{(x-c)^{n+1}}{(n+1)!} \Big|_{[a,b]}^{[a,b]}.$$

Each uses intersections of subsequent estimates and order adaptation as the RR algorithm does.

Formulation of the multipanel forms for the above types requires some care. Suppose the nodes $a = x_0 < x_1 < \dots < x_k = b$ are selected, and set $Y_i = [x_{i-1}, x_i]$ for $i = 1, 2, \dots, k$. For type II integrals, let

$$(6.16) \quad I_{Y_i} f = \int_{Y_i}^{Y_i} f(x) dx,$$

which is symmetric about 0. If $s, t \in Y_i$, then

$$(6.17) \quad \int_s^t f(x) dx \in I_{Y_i} f \subseteq \sum_{i=1}^k I_{Y_i} f.$$

If $s \in Y_i$ and $t \in Y_j$ with $i < j$, then

$$(6.18) \quad \begin{aligned} \int_s^t f(x) dx &= \int_s^{x_i} f(x) dx + \int_{x_i}^{x_{i+1}} f(x) dx + \dots + \int_{x_{j-1}}^t f(x) dx \\ &\in \int_{Y_i}^{Y_i} f(x) dx + \int_{Y_{i+1}}^{Y_{i+1}} f(x) dx + \dots + \int_{Y_j}^{Y_j} f(x) dx \\ &\subseteq \sum_{i=1}^k I_{Y_i} f. \end{aligned}$$

Hence,

$$(6.19) \quad \int_{[a,b]}^{[a,b]} f(x) dx \subseteq \sum_{i=1}^k \int_{Y_i}^{Y_i} f(x) dx.$$

The subintervals Y_i are chosen by the same subinterval adaptive strategy as used for type RR integrals. Thus, the algorithm for type II integrals is the same as for type RR, except that the integral on each subinterval is computed using the one-panel type II rule, equation (6.15), with intersection.

The multipanel rules for types RI and IR do not lend themselves to subinterval adaptation. By definition,

$$(6.20) \quad \begin{aligned} \int_a^{[a,b]} f(x) dx &= \left\{ \int_a^t f(x) dx \mid t \in [a, b] \right\} \\ &= \bigcup_{i=1}^k \left\{ \int_a^t f(x) dx \mid t \in Y_i \right\} \\ &= \int_a^{Y_1} f(x) dx + \\ &\quad + \int_a^{x_1} f(x) dx + \int_{x_1}^{Y_2} f(x) dx + \\ &\quad + \dots + \\ &\quad + \int_a^{x_1} f(x) dx + \int_{x_1}^{x_2} f(x) dx + \dots + \int_{x_{k-1}}^{Y_k} f(x) dx \end{aligned}$$

Once the endpoints of a subinterval are chosen, that subinterval can be subdivided only with great difficulty. We observe that

$$(6.21) \quad \int_a^{\{a,b\}} f(x)dx \subseteq \int_{\{a,b\}}^{\{a,b\}} f(x)dx,$$

so the type II algorithm gives bounds using adaptation. If those bounds are not small enough, then we apply the algorithm given by equation (6.20) using a fixed stepsize equal to the smallest stepsize chosen adaptively by the type II algorithm.

Type IR integrals are done similarly.

In the general case of the integral (6.3), the allowable tolerances and the maximum number of function evaluations must be apportioned among three distinct, independent integrations. These are given in proportion to the width of the respective subintervals. If $\tau(X_i)$ denotes the tolerance allowed on the subinterval X_i , τ^* is the tolerance remaining, then

$$(6.21) \quad \tau(X_i) = \tau^* \cdot \frac{w(X_i)}{w(X)}.$$

If the integral on the first subinterval is narrower than its share of the total tolerance requires, then the tolerances on the other subintervals are relaxed so that the total tolerance can be met more efficiently. On the other hand, if the integral on the first subinterval is a little too wide, then the integrals on the remaining subintervals can sometimes be commuted accurately enough that the total tolerance can still be satisfied.

7. Implementation Details. It follows from the above discussion that realization of adaptive, self-validating quadrature routines on a computer requires that the following features be supported:

- (i) Interval computation (arithmetic operations and standard library functions).
- (ii) Subroutines for automatic generation of Taylor coefficients.
- (iii) Accurate scalar product of interval vectors (to minimize width due to roundoff error).

The program INTE [9] was written in FORTRAN for the Sperry 1100, on which only (i) [25] and (ii) were supported. As mentioned above, INTE performs self-validating numerical integration without adaptation. The microcomputer language Pascal-SC [22] and the ACRITH package for the IBM 370 series of computers support (i) and (iii), to which the routines (ii) have been added. The implementation described in [6] for Pascal-SC is immediately adaptable to ACRITH. The program described here is written in FORTRAN for use with ACRITH. In particular, the following operators and library functions are

supported:

+	SIN	SINH	EXP
-	COS	COSH	LOG=LN
*	TAN	TANH	LOG10
/	COTAN=COT	COTANH=COTH	ERF
** constant	ASIN	ASINH	ERFC
ABS	ACOS	ACOSH	GAMMA
SQR	ATAN	ATANH	LGAMMA
SQRT	ACOTAN=ACOT	ACOTNH=ACOTH	

The Taylor series terms $F_n(C, H)$ and $F_n(X, H)$ are calculated by using the well-known recurrence relations [15], [16], [19] for the operators and functions given above. In this application, the "point" of expansion and the stepsize are interval-valued to give the desired inclusions, but the recurrence relations remain the same. It is possible for the user to augment the list of library functions given above if the required recurrence relation for Taylor coefficients of the new function is known.

Given an integrand of the type considered, for example,

$$(7.1) \quad f(x) = \frac{4}{1+x^2},$$

the program first parses it into a *code list* [19]:

Operator	Operand 1	Operand 2	Result
SQR	x (=Temp1)		Temp4
+	1 (=Temp2)	Temp4	Temp5
/	4 (=Temp3)	Temp5	F

In order to compute the series for f expanded at C with stepsize H , this code list is interpreted to obtain the sequence of calls

```
Temp1 := (C,H) {The series for x.}
Temp2 := (1)   {Constant series.}
Temp3 := (4)   {Constant series.}
Call ITSQR(Temp1,Temp4)
Call ITADD(Temp2,Temp4,Temp5)
Call ITDIV(Temp3,Temp5,F).
```

The result is an array which contains the interval-valued series for f and an indication of how many terms were computed. For example, the subroutine ITSQR(U, V) computes the series for $V = U^2$, given the series for U , by means of the recurrence relations

$$(7.2) \quad V_i = \begin{cases} 2 * \sum_{j=1}^{\frac{i-1}{2}} U_j * U_{i-j+1} + \text{ISQR}(U_{\frac{i+1}{2}}), & i \text{ odd,} \\ 2 * \sum_{j=1}^{\frac{i}{2}} U_j * U_{i-j+1}, & i \text{ even,} \end{cases}$$

where $V_i = V_i(C, H)$ and $U_i = U_i(C, H)$ denote the i th Taylor coefficients of U and V , respectively [19], p. 49. The interval function ISQR computes $\text{ISQR}(X) = X^2$ instead of $X \cdot X$, which is preferable in general, since, for example, $[-1, 1]^2 = [0, 1]$ while $[-1, 1] \cdot [-1, 1] = [-1, 1]$. The parsing and interpretation at runtime described above is unnecessary in Pascal-SC, because the compiler generates the required code [6].

The interval Taylor coefficients $F_1(X, C)$, $F_2(X, C)$, ... are maintained in a record-like structure. If "F" is the name of the function being expanded, then

LF	Index of last known nonzero term;
MF	Index of last known term;
OFL	Vector of series terms—left (lower) bound;
OFR	Vector of series terms—right (upper) bound.

The designations for other variables replace "F" in the above.

The algorithms used depend on whether the endpoints of the interval X of integration are elements of \mathbf{S} , that is, machine numbers, or whether they are intervals. The basic algorithm is for the case $X = [a, b] \in \mathbf{IS}$, and is called the RR (REAL-REAL) algorithm:

1. Compute the integral on $[a, b]$;
2. Add $[a, b]$ to the list of subintervals;
3. Loop
 4. Find the subinterval on which the width of the integral is largest;
 5. Bisect it;
 6. Compute the integral on the left subinterval;
 7. Add the left subinterval to the list;
 8. Compute the integral on the right subinterval;
 9. Add the right subinterval to the list;
 10. Compute the integral on $[a, b]$ by summing the integrals on all the subintervals;
 11. Exit when accuracy tolerance is met;
 12. Exit with warning when
 13. no further improvement in accuracy is possible,
 14. or M function evaluations are exceeded;
15. End loop.

Each subinterval X is maintained in a data structure of the following form:

XA	Left endpoint of the subinterval;
XB	Right endpoint of the subinterval;
OPTORD	Order of the derivative used to compute the remainder;
WIDINT	Width of the integral on this subinterval;
SINT	Interval-valued integral on this subinterval;
WEGHT	Vector of interval valued weights (Gauss and Newton-Cotes), stepsize (Taylor);
FNVAL	Vector of interval-valued function values (Gauss and Newton-Cotes), series terms (Taylor);
FNTRN	Vector of interval-valued function values (Gauss and Newton-Cotes), series terms (Taylor), including remainder terms.

At steps 1, 6, and 8, the integral is computed on the subinterval $[XA, XB]$ using one-panel versions of Gauss, Newton-Cotes, or Taylor polynomials as outlined in Sections 2 and 3. The weight vectors and functions are arranged in the vectors WEGHT and FNTRN, respectively, in such a way that the interval inclusion

$$(7.3) \quad J = \sum_i \text{WEGHT}_i * \text{FNTRN}_i$$

of $\int_a^b f(x)dx$ is computed as a single inner product. Similarly,

$$(7.4) \quad Rf = \sum_i \text{WEGHT}_i * \text{FNVAL}_i.$$

At step 13, if $J \subseteq Rf$, then the loop is exited. In this case, further reduction of the width of the truncation error cannot reduce $w(J)$. For Gauss and Newton-Cotes formulas, SINT is used only to give WIDINT. For Taylor polynomials, $\sum_i \text{SINT}_i$ is intersected with (7.3). SINT_i is computed using the intersection principle discussed at the end of §3. For a few subintervals, $\sum_i \text{SINT}_i$ is narrower than (7.3), while the situation is usually reversed for a large number of subintervals, because (7.3) uses the accurate scalar product.

The arrangement of subintervals in the arrays listed above must be relatively straightforward to allow J to be computed by a single scalar product operation. Each iteration of the loop from step 3 to step 15 removes one subinterval from the list and replaces it with two subintervals. Subintervals are not otherwise deleted from the list. Hence, the following simple allocation scheme works: On the i th pass through the loop, the information about the left subinterval is stored in the locations previously used by its parent, and the information about the right subinterval is stored in the $(i + 1)$ st locations, following the already computed values. Hence, insertion requires no searching. The widest subinterval is found at step 4 by a sequential search of the array WIDINT(1.. i).

By contrast, QUADPACK [17] maintains its list of pending subintervals in sorted order, so no search is necessary for the next subinterval to be processed. However, new subintervals are inserted at locations found by a sequential search, followed by changing pointers to all following entries in the list. For each subinterval processed, the program does one sequential search, while QUADPACK does two. In addition, QUADPACK uses two sets of pointer adjustments.

For integration using Taylor polynomials, the maintenance of list of subintervals is somewhat more complicated, because the program reuses the series which it has previously computed. To illustrate the ideas, consider the first execution of the loop at step 3. At that point, the list of subintervals contains only one: $X = [a, b]$ itself. FNTRN contains $\text{OPTORD}-1$ terms of the series for f expanded at $c = (a + b)/2$ and the truncation error term involving $F^{(\text{OPTORD})}(X)$ given by equation (4.6). Provided that the requested tolerance exceeds the noise inherent in the function evaluation, a stepsize h can be computed which is small enough that the requested tolerance per unit step is satisfied on the interval $[c - h, c + h]$. Notice that if a relative tolerance is requested, then this requires a current estimate for J . The value of the integral on this subinterval can be computed at a cost proportional to OPTORD , instead of a cost proportional to OPTORD^2 , which would be required to generate J directly by using the recurrence relations for f . Following this, the two subintervals $[a, c - h]$ and $[c + h, b]$ are processed directly. Consequently, this method breaks the subinterval of integration into three parts, rather than bisecting it.

Thus, for integration by Taylor polynomials, step 5 of the RR algorithm is replaced by

- 5.0' Compute h such that the tolerance is satisfied on $[c - h, c + h]$;
- 5.1' Compute the integral on $[c - h, c + h]$ from information in FNVAL;
- 5.2' "left subinterval" := $[XA, c - h]$;
- 5.3' "right subinterval" := $[c + h, XB]$.

The middle subinterval $[c - h, c + h]$ is maintained on the list of subintervals so that its contribution to J is included in the scalar product in (7.3). This has the helpful side effect that h can be chosen somewhat optimistically. If the choice is too optimistic, then $[c - h, c + h]$ will be selected later for further processing as the worst subinterval, at which time it will be broken into three parts. One of the new subintervals will occupy the place of the parent interval in the list maintained, while the other two will be added to the end of the list.

A further refinement could be implemented. The stepsize h computed at step 5.0' has the following property: On each subinterval of length $2h$ which is contained in $[XA, XB]$, the use of a Taylor polynomial of degree $\text{OPTORD}-1$ yields an integral which satisfies the requested tolerance. The integration on such a subinterval can be done with half the usual work. No series for the truncation error needs to be computed because the truncation error can be bounded by using the global remainder term on $[XA, XB]$. If $[XA, XB]$ can be covered by a few subintervals of length $2h$, then this could be done, and division into three parts would be needed only when the middle part is relatively small. This refinement could improve the efficiency of the program. However, the improvement would likely be modest, because the advantages of intersecting subsequent estimates on each small subinterval would be lost, and the number of subintervals added to the list would

no longer be constant. The program also does not reuse function evaluations required by Gauss or Newton-Cotes formulas, although it could be modified to do so.

8. Numerical Examples. We give four examples to illustrate the accuracy and the reliability of the program. All computations were done in double precision on an IBM 4341 computer, using calls to ACRITH routines for all necessary interval calculations, including scalar products.

$$\text{Example 1. } I = \int_{0.6}^{0.7} \frac{1}{1-x} dx.$$

Interval bounds for the answer can be computed in three ways:

$$(8.1) \quad I = \log(1 - 0.6) - \log(1 - 0.7),$$

$$(8.2) \quad I = \log(4/3),$$

or by adaptive, self-validating quadrature.

Neither 0.6 nor 0.7 are machine numbers, so they are converted to intervals which are one machine number wide. All three methods give the interval [0.2876820724517808, 0.2876820724517811], but the results are 20, 13, and 14 machine numbers wide, respectively. That is, the program is capable of accuracies comparable to evaluation of the analytic expression for the answer. This result required 43 equivalent function evaluations on 18 subintervals. In order to appreciate the accuracy achieved by the program, we must consider some details at the level of machine numbers. If (8.1) is evaluated without simplification using interval calculations, the result is the hexadecimal interval [Z 4049 A588 44D3 6E41, Z 4049 A588 44D3 6E55], which is 20 machine numbers wide. Using (8.2), the last 4 hexadecimal digits are [Z ... 6E45, Z ... 6E52], which is 13 machine numbers wide. The adaptive, self-validating quadrature program gives [Z ... 6E44, Z ... 6E52], which is 14 machine numbers wide.

$$\text{Example 2. } \int_0^4 \sqrt{x} dx = \frac{16}{3} \text{ (see §5.1).}$$

This example illustrates the order adaptation required to handle the nonexistence of $f'(0)$. The program gave the interval [5.33333 33333 27, 5.33333 33333 36]. It stopped when it had used 400 effective function evaluations on 226 subintervals. Most of the subintervals were clustered near the origin. Away from 0, as many as 13 series terms were used.

$$\text{Example 3. } \int_0^1 f(x) dx, \text{ where } f(x) = \begin{cases} 0, & x < 0.3, \\ 1 & x \geq 0.3. \end{cases}$$

This example illustrates integration of a simple discontinuous function. The parser does not accept a piece-wise definition, so this function was coded by hand. Table 1 shows the results for tolerances 0.0 and 1.0E-15, and for recognizing that the function has a singularity on the interval of integration.

	Width of Integral	Function Evaluations	Subintervals
Tolerance = 0.0			
$[0, 1]$	2.50E-16	218	146
$[0, 0.3] + [0.3, 1]$	2.78E-17	22	29
Tolerance = 1.0E-15			
$[0, 1]$	7.77E-16	194	130
$[0, 0.3] + [0.3, 1]$	1.39E-16	10	5

Table 1. Integrating a Discontinuous Function.

As is usually the case, the program performs *much* better when the user recognizes the presence of a discontinuity. The performance would be even better if the discontinuity occurred at a machine number. Notice that the cost of requesting the program to do the best it can (tolerance = 0.0) is only slightly more than the cost when a lesser tolerance is prescribed. The table shows more subintervals than evaluations because evaluating a series which is $\equiv 0$ is not counted as an evaluation.

Example 4. $\int_0^1 \frac{1}{1 - \alpha x^2} dx.$

This example illustrates the performance of the program as the integrand varies from very smooth to being undefined at a point in the interval of integration. These calculations were done in single precision with an absolute error tolerance request of 1.0E-5. An error code of 66 signals that the program was unable to meet the requested tolerance, while 67 means that it was unable to evaluate the integrand. As the problem became more difficult, the program required more effective function evaluations and more subintervals.

Alpha	Error Code	Function Evaluations	Subintervals	Maximum Order	Absolute Error
0.00	0	2	2	3	0.0
0.05	0	8	2	20	4.4E-16
0.10	0	8	2	20	1.8E-14
0.15	0	8	2	20	6.6E-12
0.20	0	8	2	20	7.8E-10
0.25	0	5	2	15	4.1E-06
0.30	0	15	6	15	1.9E-10
0.35	0	12	6	13	4.5E-08
0.40	0	12	6	13	1.8E-07
0.45	0	21	10	15	6.1E-08
0.50	0	19	10	14	1.4E-07
0.55	0	16	10	12	7.9E-06
0.60	0	24	14	12	2.4E-07
0.65	0	35	18	15	6.4E-08
0.70	0	29	18	12	5.0E-06
0.75	0	38	22	12	1.3E-07
0.80	0	48	26	15	6.6E-08
0.85	0	43	26	12	7.5E-06
0.90	0	62	34	15	6.8E-08
0.95	0	73	42	14	1.2E-07
1.00-	66	254	150	8	3.2E-01
1.05	67	2			

Table 2. Performance Profile.

References

1. G. Alefeld and J. Herzberger. Introduction to Interval Computation, tr. by J. Rokne. Academic Press, New York, 1983.
2. Carl de Boor. On writing an automatic integration algorithm, pp. 201-209 in *Mathematical Software*, ed. by John R. Rice, Academic Press, New York, 1971.
3. Carl de Boor. An algorithm for numerical quadrature, pp. 417-449 in *Mathematical Software*, ed. by John R. Rice, Academic Press, New York, 1971.
4. Ole Caprani, Kaj Madsen, and L. B. Rall. Integration of interval functions. *SIAM J. Math. Anal.* **12**, no. 3 (1981), 321-341.
5. G. F. Corliss and Y. F. Chang. Solving ordinary differential equations using Taylor series. *ACM Trans. Math. Software* **8** (1982), 114-144.
6. G. F. Corliss and L. B. Rall. Automatic generation of Taylor coefficients in Pascal-SC: Basic applications to ordinary differential equations. *Transactions of the First*

- Army Conference on Applied Mathematics and Computing*, pp. 177-209. U. S. Army Research Office, Research Triangle Park, N. C., 1984.
7. P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*, 2nd ed. Academic Press, New York, 1984.
8. Julia H. Gray and L. B. Rall. A computational system for numerical integration with rigorous error estimation. *Proceedings of the 1974 Army Numerical Analysis Conference*, pp. 341-355. U. S. Army Research Office, Research Triangle Park, N. C., 1974.
9. Julia H. Gray and L. B. Rall. INTE: A UNIVAC 1108/1110 program for numerical integration with rigorous error estimation. *MRC Technical Summary Report No. 1428*, University of Wisconsin-Madison, 1975.
10. Julia H. Gray and L. B. Rall. Automatic Euler-Maclaurin integration. *Proceedings of the 1976 Army Numerical Analysis and Computers Conference*, pp. 431-444. U. S. Army Research Office, Research Triangle Park, N. C., 1976.
11. U. W. Kulisch and W. L. Miranker. *Computer Arithmetic in Theory and Practice*. Academic Press, New York, 1981.
12. U. W. Kulisch and W. L. Miranker (Eds.). *A New Approach to Scientific Computation*. Academic Press, New York, 1983.
13. M. A. Malcolm and R. B. Simpson. Local vs. global strategies for adaptive quadrature. *ACM Trans. Math. Software* 1, no. 2 (1975), 127-146.
14. R. E. Moore. The automatic analysis and control of error in digital computation based on the use of interval numbers, pp. 61-130 in *Error in Digital Computation*, Vol. 1, ed. by L. B. Rall. Wiley, New York, 1965.
15. R. E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N. J., 1966.
16. R. E. Moore. *Techniques and Applications of Interval Analysis*. SIAM Studies in Applied Mathematics, 2, Society for Industrial and Applied Mathematics, Philadelphia, 1979.
17. R. Piessens, E. de Doncker-Kapenga, C. W. Überhuber, and D. K. Kahaner. *QUADPACK: A Subroutine Package for Automatic Integration*. Springer Series in Computational Mathematics, No. 1. Springer, New York, 1983.
18. L. B. Rall. Optimization of interval computation, pp. 489-498 in *Interval Mathematics 1980*, ed. by K. L. E. Nickel, Academic Press, New York, 1980.
19. L. B. Rall. *Automatic Differentiation: Techniques and Applications*. Lecture Notes in Computer Science, no. 120. Springer, New York, 1981.
20. L. B. Rall. Integration of interval functions II. The finite case. *SIAM J. Math. Anal.* 13, no. 4 (1982), 690-697.
21. L. B. Rall. Representations of intervals and optimal error bounds. *Math. of Comp.* 41, no. 163 (1983), 219-227.
22. L. B. Rall. An introduction to the scientific computing language Pascal-SC. *Transactions of the Second Army Conference on Applied Mathematics and Computing*, pp. 117-148. U. S. Army Research Office, Research Triangle Park, N. C., 1985.
23. J. R. Rice. A metalgorithm for adaptive quadrature. *J. ACM* 22, no. 1 (1975), 61-82.
24. A. H. Stroud and D. Secrest. *Gaussian Quadrature Formulas*. Prentice-Hall, Englewood Cliffs, N. J., 1966.
25. J. M. Yohe. The interval arithmetic package. *MRC Technical Summary Report No. 1755*, University of Wisconsin-Madison, 1977.

ASPECTS OF A HIGH LEVEL ALGORITHM FOR PROCESSING DIVERGING AND
CONVERGING BRANCH NONSERIAL DYNAMIC PROGRAMMING SYSTEMS

Augustine O. Esogbue
School of Industrial and Systems Engineering
Georgia Institute of Technology, Atlanta, Georgia 30332

and

Nazir A. Warsi
Department of Mathematical and Computer Sciences
Atlanta University, Atlanta, Georgia 30314

ABSTRACT

We report about the computational aspects of high level algorithms developed for efficiently processing the diverging and converging branch systems in nonserial dynamic programming. A special feature of these algorithms consists of a special technique devised for processing the network functions such that the minimum amount of storage is employed. It is shown that if k is the discretization level of the state and decision variables then the space complexities are $O(k)$ and $O(k^2)$ for the diverging and converging branch systems respectively. The resultant time complexities are also developed. These savings in computational complexities enhance the attractiveness of dynamic programming as a tool for processing more complex nonserial systems.

1. INTRODUCTION

When considering nonserial dynamic programming networks [12] attention is usually focussed initially on the following four basic well structured systems: diverging, converging, feed forward and feedback loop systems. A characteristic of each of these nonserial systems, is the fact that at least one subsystem either receives inputs from more than one subsystem or sends outputs to more than one subsystem. Alternatively, for at least one of the stages of these systems, the output is not the input to the next; thus, there exists at least one n such that the output $x_n \neq x_{n-1}$, the input of the next stage. This distinguishes them from the usual serial systems.

Examples of the above classical nonserial systems and various combinations of them abound in real life ([2] [13] [18]). For example, they are encountered in the study of chemical processing systems, natural gas transmission pipelines, water resources systems, energy, communication and computer networks. There are important reasons to treat these problems from the standpoint of nonserial dynamic programming. In general, however, the resultant problems are more difficult computationally than classical serial dynamic programming. It is clear that computational advances in serial dynamic programming play an important

role in the study of nonserial dynamic programming processes. As an extension of this argument, analysis of complex nonserial dynamic programming processes consisting of various combinations of each of the basic structures is also aided by computational advances in the four classical structured nonserial dynamic programming systems outlined earlier.

Hitherto, limited attention has been paid to efficient algorithms for treating nonserial dynamic programming networks. In particular there is a complete absence of any discussions relative to their computational complexities. This shortcoming was recently addressed in [14]. The emphasis was on the development of efficient computing algorithms which will minimize the usual storage requirements of dynamic programming encountered while processing the first two classical nonserial systems, namely the diverging and converging branch systems. In this paper we merely outline the problems and concentrate on the complexity analysis of the resultant algorithms. The reader is referred to the above referenced paper for algorithmic details.

2.1 THE BASIC DYNAMIC PROGRAMMING ALGORITHM FOR THE DIVERGING BRANCH SYSTEM

A diverging branch system (see Fig. 1) is the easiest of the elementary nonserial structures to analyze. For simplicity, we first consider a two branch system. The stage transformations $t(.,.)$ and return functions $r(.,.)$ both for a main serial process i ($i=1,2, \dots, n$) and for a branch j ($j = 1,2, \dots, m$) are defined as follows:

$$x_{i-1} = t_i(x_i, d_i) \quad , \quad i = 1, 2, \dots, n$$

$$x_{j-1,1} = t_{j1}(x_{j1}, d_{j1}) \quad , \quad j = 1, 2, \dots, m$$

$$x_{m1} = t_{s1}(x_s, d_s)$$

$$r_i = r_i(x_i, d_i) \quad , \quad i = 1, 2, \dots, n$$

$$r_{j,1} = r_{j,1}(x_{j1}, d_{j1}) \quad , \quad j = 1, 2, \dots, m$$

Without loss of generality, consider the basic system consisting of one main serial system ($i = 1, 2, \dots, n$) and one branch ($j = 1, 2, \dots, m$). Let us assume that the input and decision variables at each stage have the following integer values:

$$1 \leq x_{j1} \leq k_{j1} \quad , \quad j = 1, 2, \dots, m$$

$$1 \leq x_i \leq k_i \quad , \quad i = 1, 2, \dots, n$$

$$1 \leq d_{j1} \leq p_{j1} \quad , \quad j = 1, 2, \dots, m$$

$$1 \leq d_i \leq p_i \quad , \quad i = 1, 2, \dots, n$$

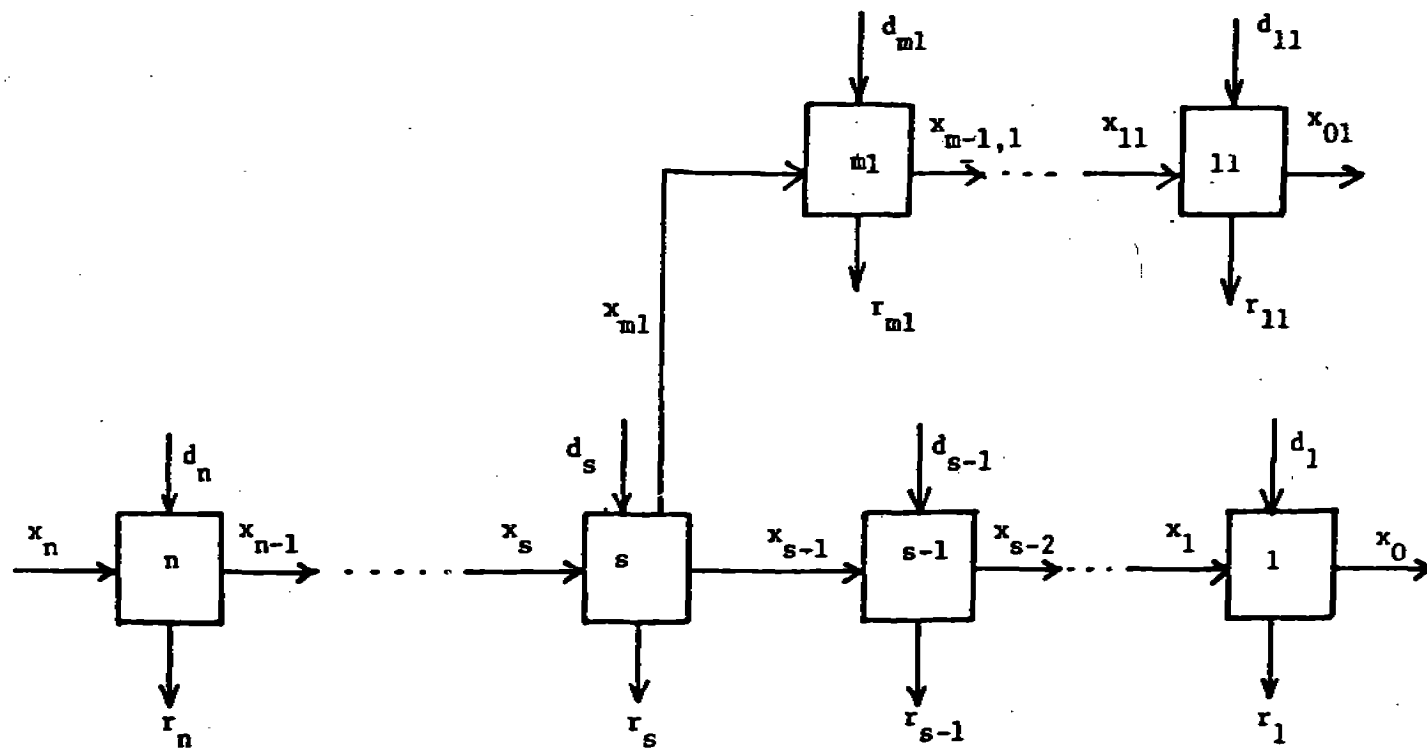


Figure 1. A Single Diverging Branch System

The algorithm is developed by first decomposing the network into four phases and then employing the usual recursive procedures in optimizing the total return. The recursion equations for the various phases then become:

2.1.1) For the Diverging Branch (from stage 11 to stage m1)

$$f_{11}(x_{11}) = \max_{1 \leq d_{11} \leq p_{11}} r_{11}(x_{11}, d_{11})$$

$$f_{j1}(x_{j1}) = \max_{1 \leq d_{j1} \leq p_{j1}} [r_{j1}(x_{j1}, d_{j1}) + f_{j-1,1}(t_{j1}(x_{j1}, d_{j1}))]$$

where $j = 2, 3, \dots$,

Using the above equations, the optimal branch return and optimal decisions are computed for each possible value of x_{i1} .

2.1.2) For the Main Serial Process (from stage 1 to stage s-1, i.e. prior to junction node)

$$f_1(x_1) = \max_{1 \leq d_1 \leq p_1} r_1(x_1, d_1)$$

$$f_i(x_i) = \max_{1 \leq d_i \leq p_i} [r_i(x_i, d_i) + f_{i-1}(t_i(x_i, d_i))]$$

where $i = 2, \dots, s-1$

The optimal return $f_i(x_i)$ and optimal decision d_i at each stage are saved for each possible input value x_i .

2.1.3) For the Stage s (junction)

$$f_{s+m1}(x_s) = \max_{1 \leq d_s \leq p_s} [r_s(x_s, d_s) + f_{s-1}(t_s(x_s, d_s)) + f_{m1}(t_{s1}(r_s, d_s))]$$

At this stage, the optimal return $f_{s+m1}(x_s)$ is the combination of main serial process preceding stage s, $f_{s-1}(t_s(x_s, d_{s1}))$ and the optimal return from the branch, $f_{m1}(t_{s1}(x_s, d_s))$. For each possible value x_s , both $f_{s+m1}(x_s)$ and d_s are reserved at this stage.

2.1.4) For the Remaining Stages (from stage $s + 1$ to stage n , the terminal node)

The optimal return at each remaining stage from $s + 1$ to n can be obtained as in the usual serial systems, i.e.,

$$f_{n+ml}(x_n) = \max_{1 \leq d_n \leq p_n} [r_n(x_n, d_n) + f_{n-1+ml}(t_n(x_n, d_n))]$$

where $i = s + 1, \dots, n$

2.1.5) Determination of the Optimal Decision and Return at Each Stage

At the final stage n , the optimal input x_n^* to the system can be obtained by letting

$$f_{n+1}(x_n^*) = \max_{1 \leq x_n \leq k_n} [f_{n+ml}(x_n)]$$

With the optimal input x_n^* and optimal d_n^* obtained from a decision table, we can produce optimal stage return r_n^* and optimal stage output x_{n-1}^* as follows:

$$r_n^* = r_n(x_n^*, d_n^*)$$

$$x_{n-1}^* = t_m(x_n^*, d_n^*)$$

This process continues from stage n down to stage $s + 1$. At the junction stage (stage s), the optimal stage input x_s^* and stage decision d_s^* , the optimal branch input x_{ml}^* are obtained via the transition equation

$$x_{ml}^* = t_{s1}(x_s^*, d_s^*).$$

For the remaining processes, the stage transformation, return function, and decision tables can be used at each stage.

2.2 A High Level Computing Algorithm for Diverging Branch Systems (DBCA)

The conventional computer algorithm for performing the operations listed in the foregoing generally employs a brute-force method to store optimal decisions at each stage and then later retracing them after the optimal return $f_n(x_n^*)$ has been found. Such an approach dictates an enormous amount of storage requirement. The problem obviously gets worse when large and complex networks are involved. This is certainly the case when multi state variable dynamic programming problems as in converging and loop systems are involved. To mitigate this problem, we have

developed a technique which marks the optimal decision value, say, d_i^* by adding $k_d = 1 + k$ to the state entry $t_i(x_i, d_i^*)$ as the processing of each stage takes place. In other words, k_d is a number larger than all discretization levels and $k = \max(k_{11}, \dots, k_{m1}; k_1, \dots, k_n)$. This eliminates the need for storing the optimal decision. Later, when the optimal decision d_i^* is to be retraced, it may be retrieved by searching only the x_i -row of t_i for $t_i(x_i, d_i^*) > k_d$ over values of d_i . This is done for each i . Any future reference to the table t_i is then made as (each entry) mod k_d .

This idea proposed for the elimination of the need for storing optimal decisions can result in a substantial saving of storage space. For example, consider a simple serial dynamic programming system involving S stages, D decision variables and P state variables. If K is the discretization level of the state variable, then using the conventional computational procedure, the storage requirement for the optimal decision is approximately SDK^P . In a very simple case where $S = 10$, $D = 1$, $K = 1000$ and $P = 2$, this saving amounts to 10^7 .

In a diverging branch system, tables $F(0;1,1:K)$ are needed for processing the optimal return at each stage. At each stage i , the optimal return is processed by using the previous stage optimal return from $F(st, \cdot)$, and stored in $F(dt, \cdot)$, where $st = (i-1) \bmod 2$, and $dt = i \bmod 2$. The optimal return f_{m1} is stored in table $FM(1:K)$. Note that st and dt are used to indicate indices of source and destination tables respectively.

To present the algorithm, we first explain the notations. The algorithms are described in a PASCAL-type construct. Comments are enclosed within $(*...*)$. Enclosure of a simple or a complex statement within a loop is effected by indenting the statement. In [14] the computer algorithm for the diverging branch system is described in detail. It consists of modules A through G each corresponding to the seven different phases of optimization.

3.1 THE BASIC DP ALGORITHM FOR THE CONVERGING BRANCH SYSTEM

Let us similarly review the converse of the diverging branch system, namely a converging branch nonserial network. In its simplest form, a number of parallel serial systems join together at a junction node and then feed their outputs to a serial system. A simple example consisting of two input parallel branches and one serial output is exhibited in Fig. 2 and is used as the leitmotif for our algorithm. In general, the converging branch system is treated as an initial final value problem (often termed a two-point boundary value problem) resulting in a two dimensional optimization problem.

We begin by noting that this system is more difficult computationally than the diverging branch system. However, as before, we may still view it as consisting basically of a serial system i , $i = 1, 2, \dots, n$ and a branch j , $j = 1, 2, \dots, m$ with convergence occurring at node (stage) s . The transformation at this stage may be written as:

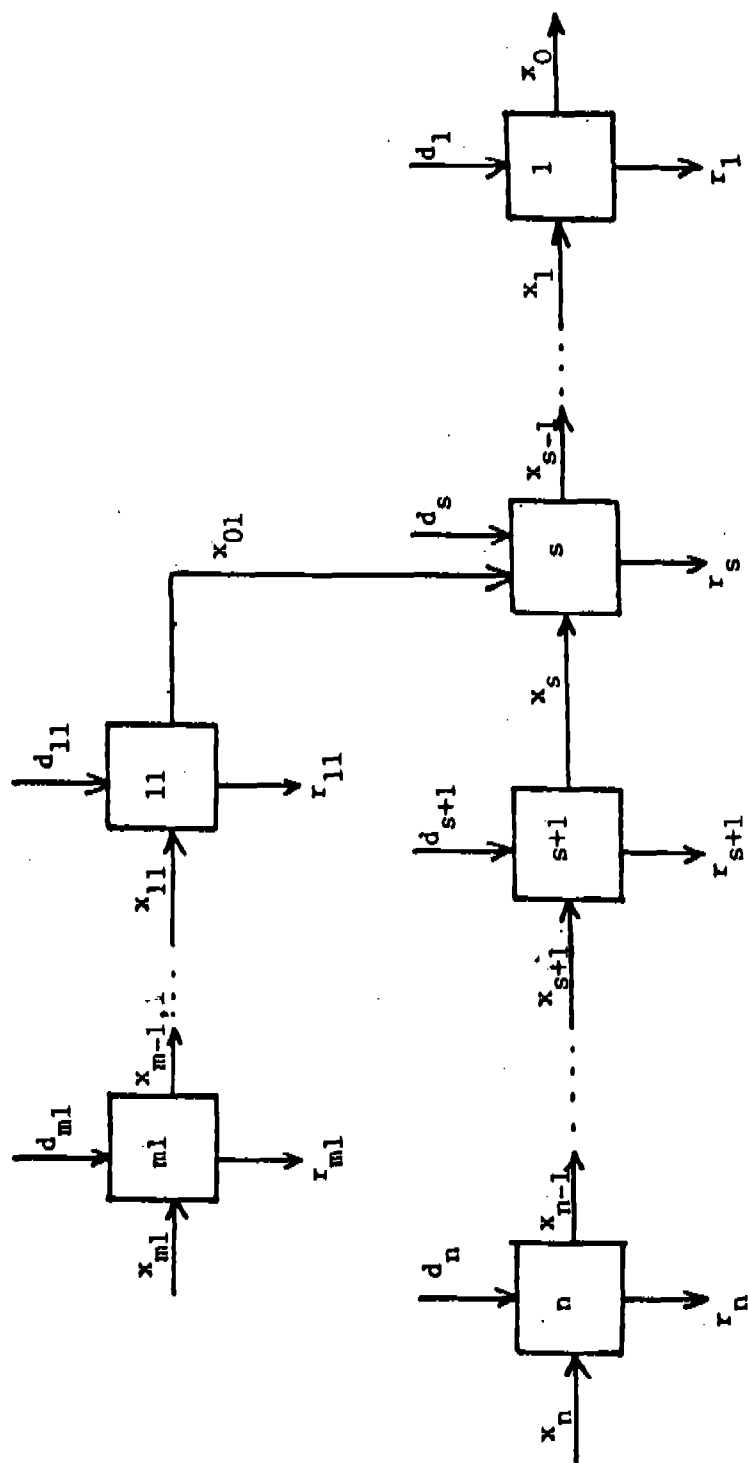


Figure 2. A Single Converging Branch System

$$x_{s-1} = t_s(x_{01}, x_s, d_s)$$

The transition function for the other stages may be represented as in the usual serial processes as follows:

For the Branches

$$x_{j-1,1} = t_{j1}(x_{j1}, d_{j1}), j = 1, 2, \dots, m$$

For the Main

$$x_{j-1} = t_j(x_j, d_j), j = 1, 2, \dots, n; j \neq s$$

We define the returns for each stage similarly. Thus,

$$r_s = r_s(x_{01}, x_s, d_s)$$

$$r_i = r_i(x_i, d_i), i = 1, 2, \dots, n; i \neq s$$

$$r_{j1} = r_{j1}(x_{j1}, d_{j1}), j = 1, 2, \dots, m$$

To develop the algorithm, we proceed as follows: We first decompose the system into three components corresponding to stages 11, 21 to m1, and 1 to n. For stages 1 to n, we separately consider stages 1 to s-1, s, and finally s+1 to n. To find the optimal branch return $f_{m1}(x_{01})$ we use the backward recursion. Next, we maximize $f_{m1}(x_{01})$ over x_{m1} . The recursion equations for the different phases may then be defined as follows:

3.1.1) FOR STAGE 11

We solve the problem

$$f_{11}(x_{11}, x_{01}) = \max_{1 \leq d_{11} \leq p_{11}} r_{11}(x_{11}, d_{11})$$

$$\text{s.t. } x_{01} = t_{11}(x_{11}, d_{11})$$

In other words, for each input value x_{11} we will find the optimal decision d_{11} which satisfies $x_{01} = t_{11}(x_{11}, d_{11})$ and also maximizes the stage return. For each value of (x_{11}, x_{01}) , the optimal decision d_{11}^* and

optimal return r_{j1} are saved. We note that a two state variable dynamic program results here as well as in the next phase of the model.

3.1.2) For Stages 21 to m1

The optimal return is given by

$$f_{j1}(x_{j1}, x_{01}) = \max_{1 \leq d_{j1} \leq p_{j1}} [r_{j1}(x_{j1}, d_{j1}) + f_{j-1,1}(t_{j1}(x_{j1}, d_{j1}))]$$

$$j = 1, 2, \dots, m$$

At each stage from 21 to m1, the optimal decision d_{j1}^* and optimal return f_{j1}^* are computed for each pair of (x_{j1}, x_{01}) . At stage m1, $f_{m1}(x_{m1}, x_{01})$ is found and the value of x_{m1}^* which maximizes the branch return for each value of x_{01} is obtained such that

$$f_{m1}(x_{01}) = \max_{x_{m1}} f_{m1}(x_{01}, x_{m1})$$

3.1.3) For the Main Serial Process

3.1.3-i The optimal return from stages 1 to $s - 1$ can be found by using the usual recursive procedure, i.e.,

$$f_1(x_1) = \max_{1 \leq d_1 \leq p_1} r_1(x_1, d_1)$$

$$f_i(x_i) = \max_{1 \leq d_i \leq p_i} [r_i(x_i, d_i) + f_{i-1}(t_i(x_i, d_i))]$$

where $i = 2, 3, \dots, s - 1$

3.1.3-ii) At Stage s (the junction node)

The optimal branch return $f_{m1}(x_{m1}, x_{01})$ is combined with the return at stage s and the optimal return from stages 1 through $s - 1$ using the recursion equation

$$f_s(x_s, x_{m1}) = \max_{1 \leq d_s \leq p_s} [r_s(x_{01}, x_s, d_s) + f_{s-1}(t_s(x_{01}, x_s, d_s))$$

$$+ f_{m1}(x_{m1}, x_{01})]$$

where the maximization is over $1 \leq x_{01} \leq k_{01}$ and $1 \leq d_s \leq p_s$. In other words, at junction s, we compute the optimal return $f_s(x_s)$ and determine optimal branch output x_{01} , and optimal decision d_s , for each input value of x_s . We can also obtain the optimal branch input x_{m1} which maximizes the branch return using the value of x_{01} . The use of the decomposition

principle at the junction stage s ensures that the optimization of the main serial chain is reduced to a sequence of one dimensional problems.

3.1.3-iii For Stage $s + 1$ to n

The recursion equation is given by

$$f_i(x_i) = \max_{1 \leq d_i \leq p_n} [r_i(x_i, d_i) + f_{i-1}(t_i(x_i, d_i))]$$

where $i = s + 1, \dots, n$

At the final stage n the optimal system return for each input value of x_i can be obtained.

3.1.4) Determination of the Optimal Decision and Return at Each Stage

At the final stage, the optimal input x_n^* can be obtained which maximizes $f_n(x_n)$ with the optimal decision d_n^* obtained from the decision table. We will proceed from stage $n - 1$ to stage $s + 1$. At stage s , the optimal input x_s^* and optimal branch input x_{01}^* are found as follows:

$$x_s^* = t_{s+1}(x_{s+1}, d_{s+1}^*)$$

Now that x_s^* has been found, the optimal branch input x_{01}^* can be obtained. This is because we decided optimal x_{01} for each value of x_s when evaluating the optimal objective function value at stage s . For the remaining stages the optimal stage input and decision can be obtained using, the stage transformation function:

$$x_n = t_{n+1}(x_{n+1}, d_{n+1}), n = s - 1, s - 2, \dots, 1$$

and the decision table, respectively.

3.1.5) Input Data Required for the Algorithm

The algorithm, akin to that developed for the diverging branch system, is designed to receive the following input specifications in Pascal:

n = the number of stages in the main serial process
 m = the number of stages in the converging branch
 s = junction stage
 k_{i1} = upperbound of the input value x_{i1} ,
 k_i = upperbound of the input value x_i ,
 p_{i1} = upperbound of the decision value d_{i1} ,
 p_i = upperbound of the decision value d_i ,

3.1.6) Output List of the Algorithm

At the completion of the operations, unless otherwise specified, the algorithm outputs are akin to those of the diverging branch algorithm.

3.2 A High Level Computing Algorithm for the Converging Branch System (CBCA)

We follow the notations of Section 2.2 and note that in processing the converging branch system, tables $F1(0:l,1:k,1:k)$ and $F(0:l,1:k)$ are needed to store the converging branch and the main branch optimal returns respectively. A table $FM(1:k)$ is also needed to store the optimal return at stage $m1$ for use later. For the details of the CBCA algorithm, see [14].

4. COMPLEXITY ANALYSIS OF THE DBCA AND CBCA COMPUTING ALGORITHMS

A major deficiency of the literature of dynamic programming algorithms is the almost complete absence and certainly inadequate treatment of their resultant complexity issues. Complexity analysis is an important component of algorithms especially when digital computation is of interest or a necessity. In this section, we discuss the complexity of the two high level computer algorithms for the diverging and converging branch nonserial systems respectively. Our analysis focuses on issues related to space (storage) and computational (time) complexities.

For storage complexity, S , we assume that each variable (computer word) takes a unit storage space. We may then calculate the demand for the storage tables during computation ignoring all input tables and intermediate variables created during the course of processing.

For a performance profile of an algorithm, we consider the computational complexity T as a function of basic operations. Apparently, comparison, mod operation, assignment and arithmetic operations may be considered as such basic operations. However, observation reveals that the total number of all of the foregoing operations performed is roughly proportional to the number of comparisons. Hence, comparison constitutes the basic operation and T of an algorithm is approximated by the number of comparisons.

As an example consider a list $[c_1, c_2, \dots, c_n]$ of values where c_i is a composition of q of these basic operations. A simple routine to find the optimum value V will be:

$$V = c_1;$$

$$J_0 = 1$$

For $j = n$ to 1 do

If $V < C_j$, then $J_0 = j$; $V = C_j$

The maximum number of comparisons in this routine is $O(n)$. Similarly, searching a value V in the list can take a maximum of one comparison. Of course, more efficient searching algorithms are available. However, for the purposes of our analysis we choose this brute-force method to get an estimate on the worst cases. We further focus on the comparison where each such operation accounts for a unit time.

A simple procedure for finding the maximum of w items taking $O(w)$ time can be constructed. A binary search taking $\log_2 w$ time may be used for searching a list of w items.

Let k be the discretization level of the state as well as the decision variables i.e. $k_{i1} = p_{i1} = k_i = p_i = k$. Also let m and n be the dimension of the stages in the branch and main subsystems respectively. We state and prove the following theorems which characterize our algorithm.

Theorem 1.

For the diverging branch computer algorithm, the space complexity $S(\text{DBS}) = O(k)$ and the time complexity, $T(\text{DBS}) = O((m+n)k^2)$

Proof

For $S(\text{DBS})$, we reason as follows:

The DBCA algorithm employs tables $F(0;1,1:k)$ and $FM(1:k)$ which consume $2k$ and k spaces respectively. Thus, $S(\text{DBS}) = 2k + k = 3k = O(k)$. For $T(\text{DBS})$ we proceed by listing the time for each of the basic computations (steps).

<u>Step</u>	<u>Time</u>
A	$\sum_{i=1}^m k_{i1} p_{i1}$
B	$\sum_{i=1}^{s-1} k_i p_i$
C	$k_s p_s$
D	$\sum_{i=s+1}^n k_i p_i$
E	k_n
F	$\sum_{i=1}^n (1 + \log p_i)$
G	$\sum_{i=1}^m \log p_i$

$$\begin{aligned}
\text{Hence, } T(\text{DBS}) &= \sum_{i=1}^m (k_{i1} p_{i1} + \log p_i) + \sum_{i=1}^n \log p_i + \sum_{i=1}^{s-1} k_i p_i \\
&\quad + k_s p_s + \sum_{i=s+1}^n k_i p_i + \sum_{i=1}^n 1 \\
&= \sum_{i=1}^m (k_{i1} p_{i1} + \log p_i) + \sum_{i=1}^n (k_i p_i + \log p_i + 1) + k_n \\
&\leq \sum_{i=1}^m (k^2 + \log k) + \sum_{i=1}^n (k^2 + \log k + 1) + k \\
&\leq (m+n) k^2 + k + (m+n) \log k + n.
\end{aligned}$$

Let α denote the right side. Then $T(\text{DBS}) \leq \alpha$. Moreover,

$$\begin{aligned}
\frac{\alpha}{(m+n)k^2} &= 1 + \left(\frac{1}{m+n}\right) \frac{1}{k} + \frac{\log k}{k^2} + \left(\frac{n}{m+n}\right) \frac{1}{k^2} \\
&= 1 + A_1 + A_2 + A_3, \text{ with the } A_i\text{'s, } i = 1, 2, 3 \text{ defined as}
\end{aligned}$$

$$A_1 = \left(\frac{1}{m+n}\right) \frac{1}{k} < \frac{1}{k} \longrightarrow 0$$

$$A_2 = \frac{\log k}{k^2} \longrightarrow 0$$

$$A_3 = \left(\frac{n}{m+n}\right) \frac{1}{k^2} < \frac{1}{k^2} \longrightarrow 0$$

Thus $\frac{\alpha}{(m+n)k^2} \longrightarrow 1$ and $T(\text{DBS}) = O(\alpha) = O((m+n)k^2)$.

Corollary 1:

The time complexity of DBS is linearly dependent on the total number of stages in the system. Further, for a fixed number of stages, $T(\text{DBS}) = O(k^2)$.

Corollary 2:

The computational complexity of DBCA is independent of the stage where the diverging branch starts.

Theorem 2:

For the converging branch computer algorithm, the space complexity $S(\text{CBS}) = O(k^2)$ while the time complexity $T(\text{CBS}) = O((m+n)k^3)$.

Proof:

The storage tables used are $F(0:1, 1:k, 1:k)$, $F(0:1, 1:k)$ and $FM(1:k)$. However, after the optimal returns from the branch have been transferred to FM , F_1 can be released and the smaller table F may be used. This means that at any one time the maximum storage used will be no more than $2k^2 + 3k$ spaces. Hence, $S(\text{CBS}) = O(k^2)$. For time complexity we list the time taken by each step as follows:

<u>Step</u>	<u>Time</u>
A	$\sum_{i=1}^m k_{i1} p_{i1} k_{01}$
B	$k_{m1} k_{01}$
C	$\sum_{i=1}^{s-1} k_i p_i$
D	$k_s p_s k_{01}$
E	$\sum_{i=s+1}^n k_i p_i$
F	k_n
G	$\sum_{i=1}^n \log p_i + \log k_{01}$
H	$\sum_{i=1}^m (k_{i1} p_{i1} + \log p_{i1}) + k_{m1}$

Thus,

$$\begin{aligned}
 T(\text{CBS}) &= \sum_{i=1}^m (\log p_{i1} + k_{i1} p_{i1} (k_{01} + 1)) + \sum_{i=1}^n (\log p_i + k_i p_i) + k_{m1} (k_{01} + 1) \\
 &\quad + k_s p_s (k_{01} - 1) + \log k_{01} + k_n \\
 &\leq \sum_{i=1}^m (\log k + k^2 (k + 1)) + \sum_{i=1}^n (\log k + k^2) + k (k + 1) + k^2 (k - 1) \\
 &\quad + \log k + k.
 \end{aligned}$$

$$= (m + 1) k^3 + (m + n) k^2 + 2k + (m + n + 1) \log k$$

$$\leq (m + n) k^3 + (m + n) k^2 + 2k + (m + n + 1) \log k$$

Let α denote the right side to that $T(\text{CBS}) \leq \alpha$. Now,

$$\frac{\alpha}{(m+n) k^3} = 1 + \frac{1}{k} + \left(\frac{2}{m+n}\right) \frac{1}{k^2} + \left(1 + \frac{1}{m+n}\right) \frac{\log k}{k^3}.$$

The second term converges to 0. The third and fourth terms are no more than $\frac{2}{k^2}$ and $\frac{2 \log k}{k^3}$ respectively. Hence each converges to 0.

Therefore, $\frac{\alpha}{(m+n) k^3} \rightarrow 1$ and $T(\text{CBS}) = O(\alpha) = O((m + n) k^3)$

Corollary 3.

$T(\text{CBS})$ is linearly dependent on the number of stages in the system. Furthermore, for a bounded number of stages $T(\text{CBS}) = O(k^3)$.

DISCUSSION

The algorithms discussed in this paper and developed extensively in [3] have been extended to other classical nonserial systems such as the feedforward and feedback loop systems. These cases are certainly more difficult than the two classical systems treated here. A natural extension of these ideas is to complex combinations of the four basic structures. Research is in progress to extend these ideas to multi-diverging and multiconverging branch systems including various forms of branching, converging and looping. Although the pattern of analysis in such systems is akin to the ones presented here, it will be shown that memory requirements are functions of the complexity of branching present in the network.

ACKNOWLEDGMENT

This work was supported in part by the Army Research Office under ARO Grant No. DAAG 29-80-G-0010 to Atlanta University and sub-contract to the Georgia Tech Research Institute as GTRI Project Nos. E-24-623 and E-24-645. Aspects of this research were presented at the Third Army Conference on Applied Mathematics and Computing held at the Georgia Institute of Technology, Atlanta, Georgia.

REFERENCES

1. Beightler, C.S., D.B. Johnson and D.J. Wilde, "Superposition In Branching Allocation Problems", Journal of Mathematical Analysis and Applications, Vol. 12, 1965, pp. 65-70.
2. Beightler, C.S. and William Meier, "Design of Optimum Branched Allocation System," Industrial and Engineering Chemistry, Vol. 60, No. 2, February 1968, pp. 45-49.
3. Beightler, C.S. and William Meier, "Branch Compression and Absorption in Nonserial Multistage Systems," Journal of Mathematical Analysis and Applications, Vol. 21, 1968, pp. 426-430.
4. Bellman, R.E., A.O. Esogbue, and I. Nabeshima, Mathematical Aspects of Scheduling and Applications, Pergamon Press, 1982
5. Bertele, Umberto and Francesco Brioschi, "A New Algorithm for the Solution of the Secondary Optimization Problem in Nonserial Dynamic Programming," Journal of Mathematical Analysis and Applications, Vol. 27, 1969, pp. 565-574.
6. Bertele, Umberto and Francesco Brioschi, "A Contribution to Nonserial Dynamic Programming," Journal of Mathematical Analysis and Applications, Vol. 28, 1970, pp. 313-325.
7. Bertele, Umberto and Francesco Brioschi, "A Theorem in Nonserial Dynamic Programming," Journal of Mathematical Analysis and Applications, Vol. 29, 1970, pp. 351-353.
8. Bertele, Umberto and Francesco Brioschi, Nonserial Dynamic Programming, Academic Press, New York, 1973.
9. Brown, L.G., "Optimization of Nonserial Stochastic Decision Process by Dynamic Programming," Ph.D. Dissertation, University of Arkansas, 1971.
10. Esogbue, A.O., "Dynamic Programming Algorithms and Analysis for Nonserial Networks" Completion Report, GTRI Project Nos. E-24-623 and 645, Georgia Institute of Technology, Atlanta, Georgia, January 1983.
11. Esogbue, A.O. and Marks, Barry, "The Status of Nonserial Dynamic Programming," Management Science Theory, November, 1972, pp. 350-352.
12. Esogbue, A.O. and Marks, Barry, "Nonserial Dynamic Programming - A Survey," Operational Research Quarterly, Vol. 25, No. 2, 1974.
13. Esogbue, A.O. and Marks, Barry, "Dynamic Programming Models of the Nonserial Critical Path-Cost Problem," Management Science, Vol. 24, No. 2, 1977, pp. 200-209.
14. Esogbue, A.O. and Nazir Warsi, "A High Level Computing Algorithm for Diverging and Converging Branch Nonserial Dynamic Programming", International Journal of Computers and Mathematics. In Press.

References (Cont'd)

15. Garey, M.R. and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman and Company, San Francisco, 1979.
16. Parker, M.W., "Nonserial Multistage Systems-Analysis and Applications," Unpublished, Ph.D. Dissertation, University of Arkansas, 1969.
17. Parker, M.W. and R.M. Crisp, "Decomposition of Converging Branch Multistage Systems," AIIE Transactions, Vol 2, 1970, pp. 185-190.
18. Wong, Peter and Robert Larson, "Optimization of Tree-Structured Natural Gas Transmission Networks," Journal of Mathematical Analysis and Applications, Vol. 24, 1968, pp. 613-626.

A MODEL FOR SYMMETRIC VORTEX-MERGER[#]

M. V. MELANDER, N. J. ZABUSKY AND J. C. MCWILLIAMS*

Institute for Computational Mathematics and Applications
Department of Mathematics and Statistics
University of Pittsburgh
Pittsburgh, Pa 15260

* National Center for Atmospheric Research
P.O. Box 3000 Boulder, Colorado 80307-3000

We examine the pairing or merger of two identical regions of uniform vorticity using an approximation to the 2D-Euler equations based on second order local physical space moments. This approximation, that can describe the initial phase of a merger, yields a Hamiltonian system of ordinary differential equations for the evolution of the centroid position, aspect ratio and orientation of each region. The symmetry of the problem makes this system integrable. Thereby we obtain a necessary and sufficient condition for merger. This condition involves only the initial conditions and the conserved quantities. The existence of "pulsating" solutions, observed computationally, is related to the existence of two steady corotating states - two fixed points in a phase plane.

1. Introduction

When two vortices with like signed vorticity are sufficiently close together they merge or pair into one vortex. For example, two identical circular vortices of uniform vorticity merge when the initial centroid separation is smaller than 1.7 diameters. In spite of the importance of the merger process - especially for two-dimensional turbulence calculations (McWilliams (1985)) - there has so far been only a limited understanding of why the process takes place. In this paper we present the *first* simple analytical model of symmetric merger of uniform vortices.

Melander, Zabusky and Styczek (1985) have developed a physical-space moment description for vortex interactions of the 2D-Euler equations. Although this model is asymptotic and becomes increasingly invalid during a merger, it yields a fair description of the initial evolution towards vortex pairing. The moment model yields a Hamiltonian system with four ordinary differential equations for each vortex. When the model is applied to the present problem we

[#] Presented at the Third Army Conference on Applied Mathematics and Computing, May 1985. To be published in proceedings of the meeting.

Acknowledgement. This work was supported by the ARMY Research Office and the Office of Naval Research.

obtain an integrable model. Using this model we obtain a simple explicit formula (28) for the threshold of the process.

In Section 2, we discuss briefly the ideas behind the moment model and state the equations governing the interaction of two identical vortices. The solution to these equations is presented in Section 3 and interpreted through phase portraits in Section 4. Dynamical results derived from the moment model are validated qualitatively in Section 5 by comparing with high resolution spectral simulations.

2. Formulation of the model equations

We consider two identical regions D_1 and D_2 of *constant* vorticity ω . The regions are symmetrically situated around the global vorticity centroid, which we for convenience take as the origin of a cartesian coordinate system. The dynamics is governed by the 2D-Euler equations, which in vorticity-streamfunction form are

$$d_t \omega = \omega_t + \omega_x \psi_y - \omega_y \psi_x = 0, \quad (1)$$

and

$$\Delta \psi = -\omega. \quad (2)$$

We approximate this problem using the moment model derived in Melander, Zabusky and Styczek (1985). For the sake of the completeness we briefly describe the ideas behind this model.

The system of equations (2) and (1) has the following weak formulation

$$\psi(x) = - (1/2\pi) \int \omega(x') \ln|x-x'| d\rho', \quad (3)$$

$$d_t \int F(x) \omega d\rho = \int \omega (\nabla F) \cdot d_t x d\rho, \quad (4)$$

where $d\rho \equiv dx dy$ and F is an arbitrary test function. Assuming that D_1 and D_2 are disjoint ellipses (of common area A , aspect ratio λ and orientation ϕ as shown in Figure 1), we may express the streamfunction ψ in terms of the local moments

$$J_k^{mn} = \int_{D_k} \omega (x-x_k)^m (y-y_k)^n d\rho, \quad (5)$$

where $x_k = (x_k, y_k)$ denotes the centroid of D_k . If $x \in D_1$ then

$$\frac{1}{2\pi} \int_{D_2} \ln |x-x'| d\rho' = \sum_{m+n=0}^{\infty} C_{mn} (x-x_2)^{mn} \quad (6)$$

is a convergent series representing the far-field, while the near-field of ψ is obtained from the properties of Kirchhoff's elliptical vortex. By truncating the expansion (5) after $m+n=2$ and restricting the test functions F to be quadratic polynomials of x and y inside D_1 and D_2 , we obtain a finite system of ordinary differential equations for the local moments and the centroid positions. Using the terminology introduced in Figure 1, these equations become:

$$d_t \lambda = \frac{\lambda \bar{\omega} A \sin 2(\theta-\phi)}{4\pi R^2}, \quad (7)$$

$$d_t \phi = \frac{\bar{\omega} \lambda}{(1+\lambda)^2} + \frac{\bar{\omega} A(1+\lambda^2)}{8\pi R^2(1-\lambda^2)} \cos 2(\theta-\phi), \quad (8)$$

$$d_t \theta = \frac{A \bar{\omega}}{4\pi R^2} \left[1 - \frac{A(1-\lambda^2)}{8\pi R^2 \lambda} \cos 2(\theta-\phi) \right], \quad (9)$$

$$d_t R = \frac{A^2 \bar{\omega} (1-\lambda^2) \sin 2(\theta-\phi)}{32\pi \lambda R^3}, \quad (10)$$

where A is conserved. These equations are derived in Melander, Zabusky and Styczek (1985). Note that θ and ϕ enter only in the combination $\theta-\phi$. The model is asymptotically correct to order $O(\epsilon^3)$ where

$$\epsilon \equiv \max (\text{diameter } D_1)/(2R). \quad (11)$$

Therefore as a merger progresses the model becomes increasingly invalid. However, from past experience the model's results compare favorably with results for the full Euler equations. That is we have found a good agreement for critical merger distances - in general the deviation from contour dynamical results is only 5-10%. For example the critical merger distance for two identical circular vortices is $3.4\sqrt{A/\pi}$ for contour dynamics and $3.2\sqrt{A/\pi}$ for the moment model.

Equations (7) through (10) conserve the total angular impulse or moment of inertia

$$M = 2A\omega [R^2 + A(1+\lambda^2)/(4\pi\lambda)], \quad (12)$$

and the excess energy

$$H = 1/2 \int \omega \psi d\rho. \quad (13)$$

The first follows by integration of (7) and (10). The second follows from the truncation of the expanded streamfunction.

The system (7)–(10) becomes Hamiltonian when canonical variables are introduced. However, we shall not use these variables in this paper.

We make the equations dimensionless by introducing the following scaling

$$t = t\bar{\omega}, \quad R = R(\pi/A)^{1/2}. \quad (14)$$

The dimensionless centroid separation s becomes $2R(\pi/A)^{1/2}$, and we introduce

$$K = 1/s^2 = A/(4\pi R^2). \quad (15)$$

A convenient normalization of M is

$$\sigma = 2\pi M/\bar{\omega}A^2 > 2, \quad (16)$$

which may be written as

$$1/K = \sigma - (1+\lambda^2)/\lambda. \quad (17)$$

If $\xi \equiv \phi - \theta$ then we can rewrite the system as two first order differential equations

$$d_t \lambda = -\lambda K \sin 2\xi, \quad (18)$$

$$d_t \xi = \Omega_s - \Omega_m, \quad (19)$$

where

$$\Omega_s \equiv \lambda/(\lambda+1)^2, \quad (20)$$

$$\Omega_m \equiv (K/2) \left\{ \left[\frac{\lambda^2+1}{\lambda^2-1} + \frac{K(\lambda^2-1)}{\lambda} \right] \cos 2\xi + 2 \right\}.$$

Here Ω_s and Ω_m are the angular velocities due to the self-interaction and the mutual-interaction, respectively.

3. The solution of the model equations

With (17) inserted into (18) and (19) we obtain an autonomous system. Hence for a fixed value of σ , a conserved quantity, the problem is within the framework of a classical phase plane analysis.

Obviously there is a singularity in the equations for $\lambda=1$, which is due to the way we describe the ellipse, namely by an aspect ratio and an orientation. (Clearly the orientation is not well-defined for a circle). There are other bad features of the (λ, ϕ) -description as well. For example the same ellipse can be described in many ways, (λ, ϕ) , $(1/\lambda, \phi + \pi/2)$, etc. A convenient set of coordinates for the problem is

$$(D, G) = \frac{(\lambda^2 - 1)}{\lambda} (\cos 2\xi, \sin 2\xi) \quad (21)$$

whereby each ellipse is uniquely represented. Furthermore, the singularity at $\lambda=1$ disappears when the equations are restated in the new variables

$$1/K = \sigma - \sqrt{4 + D^2 + G^2}, \quad (22)$$

$$\begin{pmatrix} \dot{D} \\ \dot{G} \end{pmatrix} = \left(K(2+KD) - \frac{2}{2+\sqrt{4+D^2+G^2}} \right) \begin{pmatrix} G \\ -D \end{pmatrix} - \begin{pmatrix} 0 \\ K\sqrt{4+D^2+G^2} \end{pmatrix}. \quad (23)$$

In terms of the new variables (D, G) the excess energy is

$$H = DK/2 - \ln((2 + \sqrt{4+D^2+G^2})/K). \quad (24)$$

The trajectories in the (D, G) -plane are the level curves of H .

4. The merger condition

The natural way to interpret the results is by looking at the (D, G) phase plane. There is a circle of radius $(\sigma^2 - 4)^{1/2}$ and center $(0, 0)$ in the phase plane where $(d_t D)^2 + (d_t G)^2$ is infinite.

Along this circle the centroid separation vanishes, $K^{-1} = 0$. Since $R=0$ corresponds to centroid "collapse", we name this circle the collapse circle. Only that part of the phaseplane, which is inside the collapse circle is of physical interest. The moment model is valid within a certain disk centered at $(0, 0)$ and with a radius, ≈ 3.75 corresponding to an aspect ratio $\lambda=4$. Equations (23) and (24) are reflection symmetric about the D -axis, corresponding to time reversal invariance. Furthermore, we find that the aspect ratio decreases when $G < 0$ and increases when $G > 0$.

The steady state solutions of (23) have been reported in Melander, Zabusky and Styczek (1985). The aspect ratios of the steady states are the real roots of a fifth order algebraic equation

$$4\lambda^5 - 7\sigma\lambda^4 + (2\sigma^2 + \sigma + 6)\lambda^3 - (2\sigma^2 + 3\sigma)\lambda^2 + 5\sigma\lambda - 8 = 0. \quad (25)$$

(In the previous work this equation was expressed in terms of the dimensionless centroid separation $\mu = 2R(\pi/A)^{1/2}$). Figure 2 shows the reciprocal aspect ratio λ^{-1} , of the states as a function of σ^{-1} . There are one, two or three steady states depending on whether $\sigma < \sigma_{cr} \approx 11.4$, $\sigma = \sigma_{cr}$ or $\sigma > \sigma_{cr}$. All of them are located on the D-axis. A local analysis shows that B is a saddle point and A is a center. C corresponds to a solution outside the collapse circle, that is without physical meaning. Let us consider the three cases.

(A) $\sigma < \sigma_{cr}$. Since there are no steady states, therefore all initial conditions lead to centroid collapse or merger. A typical phase portrait is shown in Figure 3. The two apparent stationary points on the collapse circle show that centroid collapse always takes place at an angle of 45° in the corotating frame. However, these points are far outside the region where the model is valid.

(B) $\sigma = \sigma_{cr}$. In this case we have one unstable steady state. All other initial conditions lead to merger. We remark that the fixed point does not correspond to the limiting "figure eight" state of Saffman and Szeto (1980).

(C) $\sigma > \sigma_{cr}$. This is the most complicated and interesting case. When σ is only moderately larger than σ_{cr} , we have two physically significant stationary points. We observe from the phase portrait shown in Figure 4 that there is a critical separatrix S starting and ending at saddlepoint B. All initial conditions outside S lead to centroid collapse. Inside S we have closed orbits and the center A. Along a closed orbit surrounding the origin, ξ increases steadily, so that vortices rotate counter clockwise in the corotating frame. However, orbits not surrounding the origin, ξ oscillates around zero, corresponding to a nutation of ellipses in the corotating frame. Since H has a constant value on each trajectory in the phase plane, we can characterize S by the corresponding value of H at point B, namely

$$H_S(\sigma) = (\lambda^2 - 1) / [2(\sigma\lambda - \lambda^2 + 1)] \quad (26)$$

$$- \ln(2 + \sqrt{\lambda^2 + 1}) / \lambda$$

$$+ \ln(\sigma - (\lambda^2 + 1) / \lambda) \quad \lambda = \lambda_B$$

where λ_B denotes the aspect ratio of the unstable steady state B. The region of the phase plane inside S is then characterized by

$$H(D, G; \sigma) < H_S(\sigma) \quad \text{and} \quad \lambda < \lambda_B(\sigma). \quad (27)$$

(Note, $H < H_S(\sigma)$ between the outer separatrices and to right of B). Hence, merger or centroid collapse will occur if and only if one of the following conditions is satisfied:

$$\sigma < \sigma_{cr}; \quad (28)$$

$$H > H_S(\sigma);$$

$$\lambda > \lambda_B(\sigma).$$

This is the *merger condition* of the moment model for symmetric merger.

Now that we know the phase portraits we shall address the physical question of why merger occurs in the moment model. Consider the case $\sigma > \sigma_{cr}$. The moment model gives reasonable results only when λ is moderate, say less than four. Herewith we have reduced the interesting part of the phase plane to a disk centered at the origin. All trajectories inside this disk are either clearly inside the critical separatrix S or they come near the saddle point B . A local analysis near the saddlepoint is therefore justified. Note that a state in the neighbourhood of B is easily to recognized from the vorticity distribution, because the ellipses are aligned ($\xi=0$) and fairly elongated $\lambda > \min_{\sigma} \{\lambda_B\} \approx 2.36$. The separatrices divide this region of the phase plane into four sectors, labelled in Figure 4. Table 1 presents simple rules for determining the long time evolution (nonlinear stability) from a short time evolution (linear stability of the saddle point B). That is one sees the long time evolution from less than half a revolution in the corotating frame.

Table 1

Sector	Characteristic	Long-time evolution
1	$d_t \xi > 0$	pulsation
2	$d_{tt} \xi < 0, d_t \xi \approx 0$	merger
3	$d_t \xi < 0$	merger
4	$d_{tt} \xi > 0, d_t \xi \approx 0$	merger after one pulsation

From equation (19), we observe that $d\xi/dt$ becomes negative because the clockwise rotation caused by the mutual interaction dominates the counter-clockwise rotation of the Kirchhoff ellipse. In Melander, Zabusky and McWilliams (1985) we discuss the consequences of $\Omega_m > \Omega_s$ for the full Euler equations. Especially we show that $\Omega_m > \Omega_s$ can cause filamentation and axisymmetrization.

5. Qualitative comparison with high resolution simulations

In this section we discuss spectral simulations related to the critical separatrix S in the phaseplane. We do this to validate the conclusions obtained from the moment model and to highlight the importance of the separatrix S as a threshold to merger.

The simulations are made with a pseudospectral code (Haidvogel (1985)) which solves

$$\partial_t \omega + \partial_x \psi \partial_y \omega - \partial_y \psi \partial_x \omega = -\nu_4 \Delta^2 \omega \quad (29)$$

$$\omega = -\Delta \psi \quad (30)$$

in the periodic domain $[-\pi, \pi] \times [-\pi, \pi]$.

We specify the initial vorticity distribution $\omega(r, \phi, 0)$ as an idealized smooth distribution of compact support, and with equivorticity lines that are concentric ellipses of a common orientation and aspect ratio. Outside the ellipse $r=R_0(\phi)$ there is no vorticity and inside $r=R_1(\phi)$ the vorticity is uniform $\omega=\omega_p$. The relative steepness of the vorticity gradient is controlled by the parameter $\delta=(R_0-R_1)/R_0$. We refer to such a vorticity distribution as $V(\delta, \omega_p, a, b)$, where a and b are the major and minor axis of the outermost ellipse $r=R_0(\phi)$. We have found it convenient to specify $\omega(r, \phi, 0)$ as a distribution with a monotone profile function $f(r)$, $r>0$

$$\omega(r, \phi) = \omega_p \begin{cases} 1, & r < R_1 \\ 1 - f[(r-R_1)/(R_0-R_1)], & R_1 < r < R_0 \\ 0, & R_0 < r. \end{cases} \quad (31)$$

We select our profile function f from the one parameter family $\{f_k; k>0\}$ where

$$f_k(r) = \exp(-k/r \exp(1/(r-1))), \quad 0 < r < 1. \quad (32)$$

This function smoothly connects levels 0 and 1 at $r=0$ and $r=1$, and all its derivatives vanishes at these points. A suitable k is obtained from the natural requirement that $f(0.5)=0.5$, this implies $k=2.56085$. With this choice of k we find $f'(0.5)=\ln 8 \approx 2.08$, and approximately 90% of the variation of the function f occurs within the interval $[0.25, 0.75]$.

Figure 5 shows a numerical simulation of the evolution of two vortex regions as calculated on a 256^2 -mesh. The initial conditions are small- δ near-top-hats of elliptical shape - $V(0.2, 10., 0.6255, 0.4)$ - corresponding nearly to a corotating steady state solution Overman and Zabusky (1982) for $\sigma=12.66$. By trial and error we adjusted the centroid separation to 1.44 which is slightly smaller than the critical merger separation.

The two vortices approach and at $t=1.0$ the lower contours have joined. We observe that the state barely changes its shape in the time interval from $t=1.0$ to $t=3.0$. During this period the state looks like the unstable steady state (B) predicted by the moment model. Note the slow clockwise rotation of the individual vorticities with respect to a corotating frame (from $t=1.0$ to $t=2.5$). After $t=2.5$ we can no longer distinguish the individual vortices, since the dissipation has erased the gradient between them.

For $t > 3.0$ the compound core begins to axisymmetrize. We display pictures of this evolution because they constitute a unique example of the evolution after a merger of nearly uniform vortices (comparable with contour dynamics) and highlights the complicated entanglement of long filaments. In fact this simulation goes beyond what has been calculated with contour dynamics. At $t = 5.0$ and $t = 6.0$ we see a significant vorticity shedding and the formation of strong filaments. The almost complete absence of the "roll-up" phenomenon is due to the smooth vorticity distribution and the dissipation, but has nothing to do with the periodic boundary conditions (actually the periodic boundary conditions are in favour of the roll up). At $t = 8.0$ the vorticity shedding has stopped and a stable near elliptical core has formed. However, the presence of the filaments has a dramatic influence on the core, particularly when they reattach to the core. The figures of the evolution at $t = 9.0$, $t = 9.5$ and $t = 10.0$ represent our cleanest example of a reattachment. The evolution beyond $t = 10.0$ has not been calculated, but is within the framework of the axisymmetrization of a single vortex (Melander, McWilliams and Zabusky (1985)). We expect relaxation to axisymmetry through a couple of further breakings.

While Figure 5 corresponds to a trajectory slightly outside the critical separatrix S in the phaseplane, Figure 6 corresponds to a trajectory slightly inside S . This figure is the result of a 128 -mesh calculation. We start out with the same initial conditions as before only the centroid separation is now increased to 1.45. Except for the obvious effect of the dissipation on the vorticity gradients the evolution until $t = 2.0$ is the same as before. Later it becomes clear that the vortices rotate counter clockwise in the corotating frame, compare $t = 5.0$ and $t = 6.0$. In the time interval from $t = 2.0$ to $t = 5.0$ we observe an almost steady state corresponding to the saddlepoint in the phase plane. At $t = 7.0$ this state begins to break into two vortices and at $t = 8.5$ only the two lowest contours are joined. The vortices approach again at $t = 9.0$ and at $t = 10.0$ we observe a near recurrence to the steady state.

6. Conclusion

We have presented an integrable model governing the evolution of two identical vortices. The model yields an explicit merger condition (28), which involves both conserved quantities and initial conditions. For the most frequently occurring initial conditions - namely almost axisymmetric vortices - only the conserved quantities are needed. The threshold to merger in the moment model is a separatrix in a phase plane. Using specially chosen initial conditions for high resolution spectral simulations we have demonstrated a similar phenomenon for the full Euler equations. Especially we have shown that the threshold to merger is a balance between mutual- and self-interaction of the vortices. This provides us with a rule of thumb: *If the vortices - when they are aligned - have an aspect ratio larger than 2.3 and rotate clockwise in the corotating frame then merger will occur - otherwise not.*

REFERENCES

- Haidvogel, D. B. 1985 Particle dispersion and Lagrangian vorticity conservation in model of β -plane turbulence. J. Phys. Oceanogr., submitted
- McWilliams, J.C. 1985 The emergence of isolated vortices in turbulent flow. J. Fluid Mech. 146, 21-43.
- Melander, M.V, McWilliams, J.C. and Zabusky, N.J. 1985 Axisymmetrization and vorticity-gradient intensification of an isolated 2D-vortex. J. Fluid Mech., submitted.
- Melander, M.V, Zabusky, N.J. and McWilliams, J.C. 1985 Manuscript in preparation.
- Melander, M.V, Zabusky, N.J. and Styczek, A.S. 1985 A moment model for vortex interactions of the two-dimensional Euler equations. I. Computational validation of a Hamiltonian elliptical representation. J. Fluid Mech., submitted.
- Overman, E.A. and Zabusky, N.J. 1982 Evolution and merger of isolated vortex structures, Phys. Fluids 25, 1297-1305.
- Saffman, P. G. and Szeto, R. 1980 Equilibrium shapes of a pair of equal uniform vortices, Phys. Fluids 23, 2339-2342.
- Wu, H.M., Overman, E.A. and Zabusky, N.J. 1984 Steady state solutions of the Euler equations in two dimensions. Rotating and translating V-states with limiting cases. I. Numerical results, J.comput. Phys. 53, 42-71.

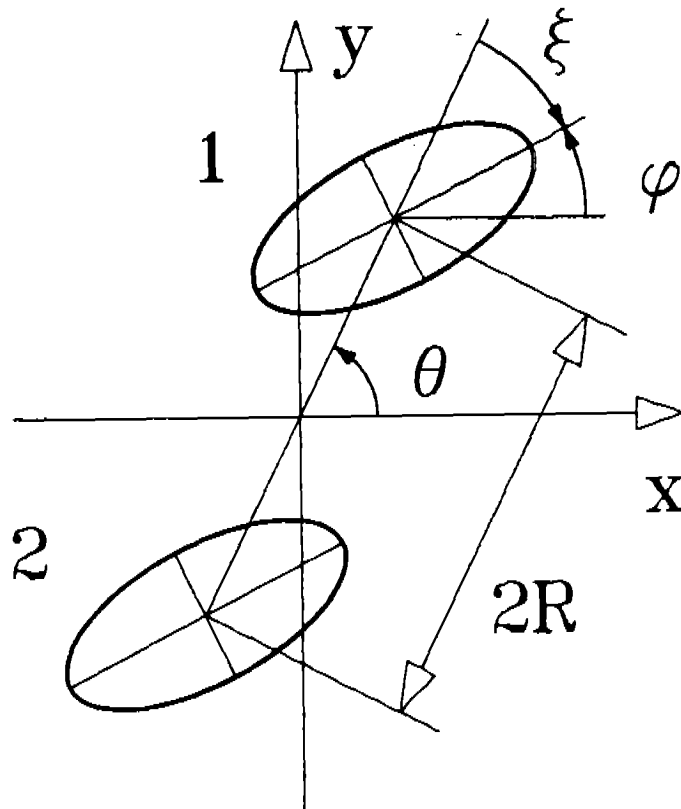


Figure 1 shows two elliptical vortices symmetrically situated around the common vorticity centroid. The centroid of vortex 1 has the polar coordinates (R, θ) and the major axis is tilted the angle ϕ . In a frame rotating with angular velocity $d_t \theta$ the orientation of the major axis is ξ .

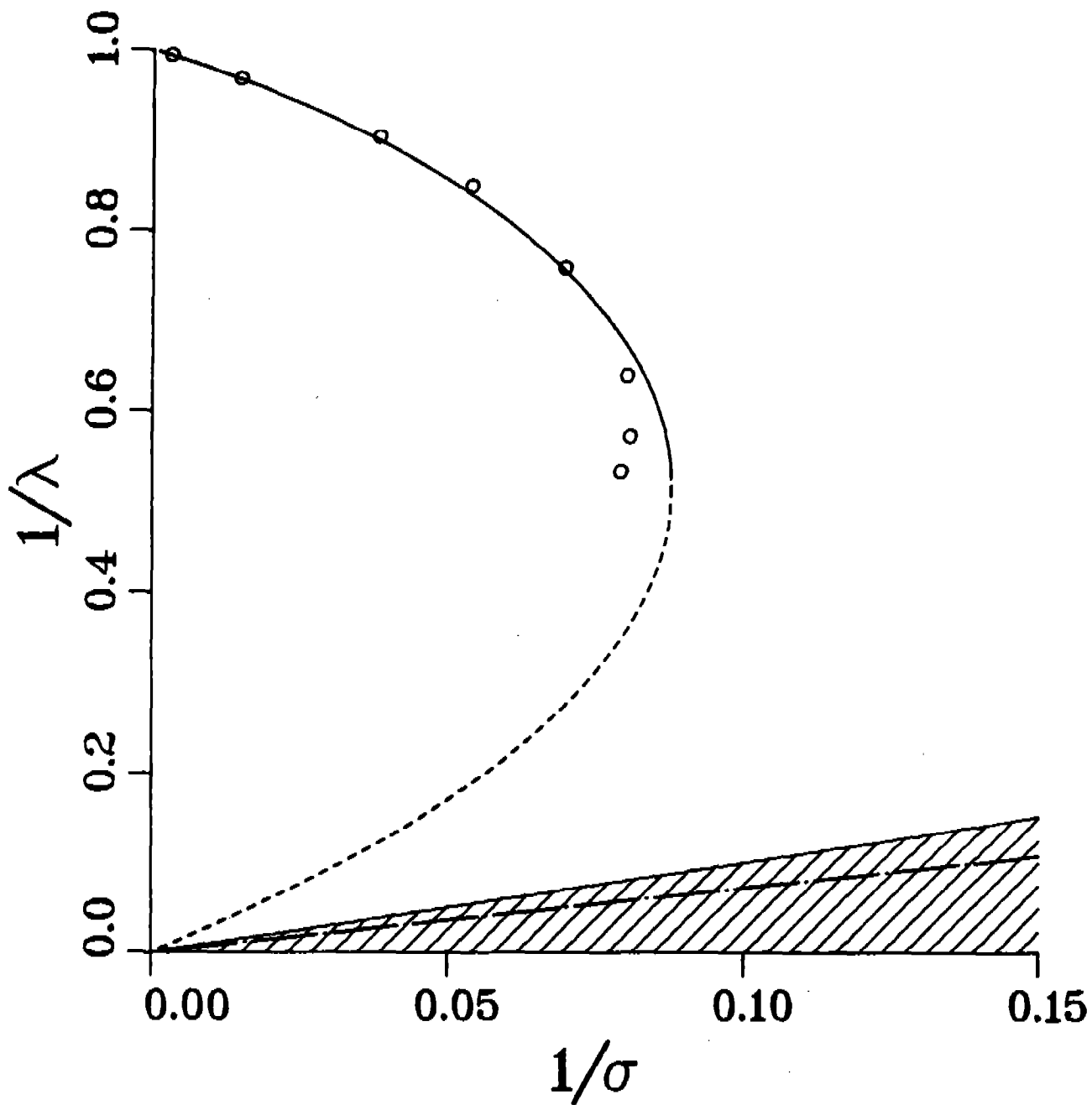


Figure 2 The aspect ratio of the steady corotating states of the moment model is plotted versus the dimensionless conserved quantity $\sigma = 2\pi M / \omega A$. A(—) are centers and B(---) are saddlepoints. For these states (A and B) the major axis of the vortices are aligned. For a given value of σ the aspect ratio λ must be smaller than $[\sigma + \sqrt{\sigma^2 - 4}] / 2$ in order to have a non-vanishing centroid separation, the unphysical region where this condition is not satisfied is shaded in the figure. The polynomial (25) has one real root C (-.-) in this region. The figure also shows results (o) from the contour dynamical calculations of Overman and Zabusky (1982).

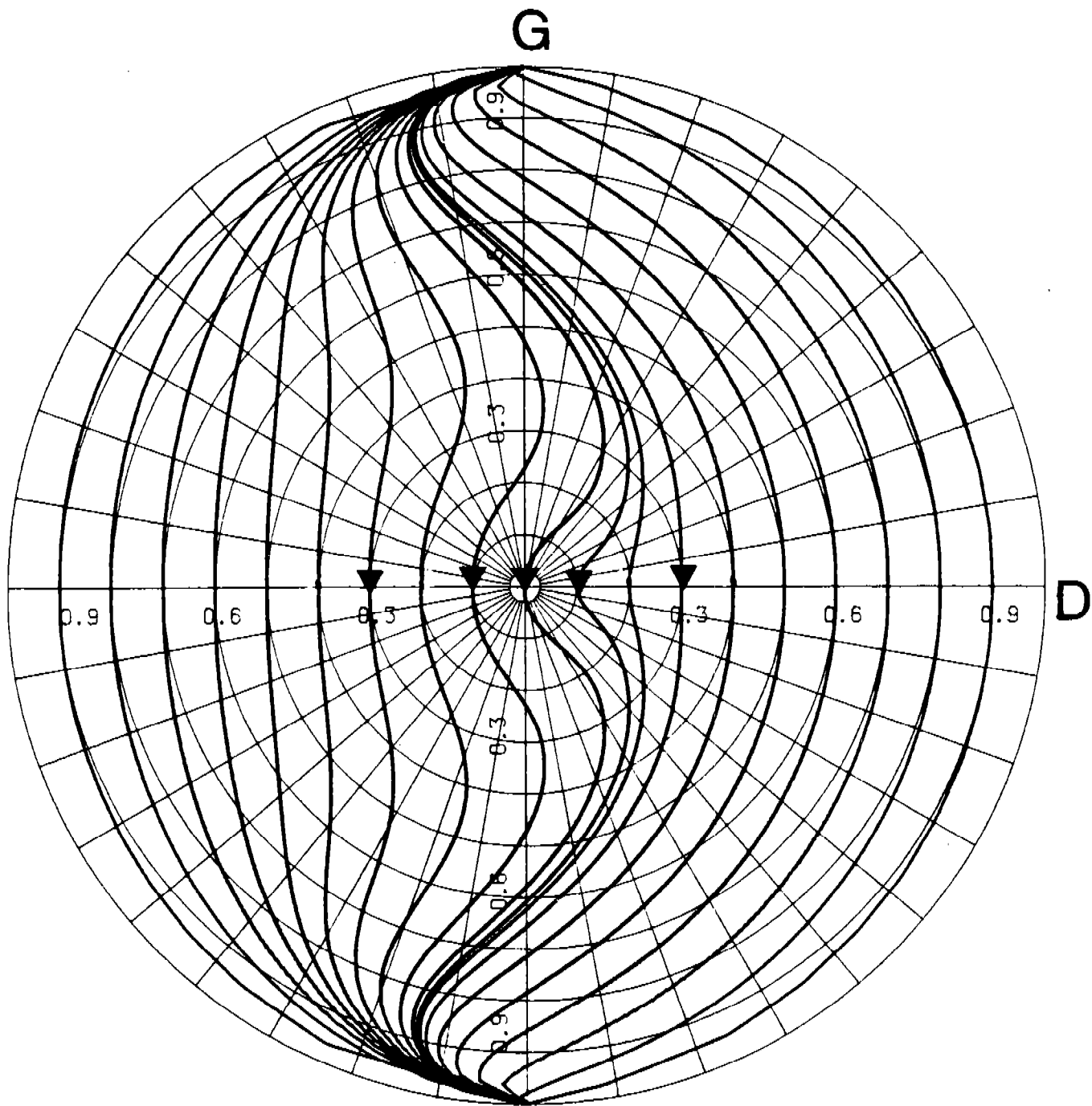


Figure 3 The phase plane for $\sigma=10.0 < \sigma_{cr}$. We have scaled the axis by the collapse radius $\sqrt{\sigma^2-4}=9.79$.

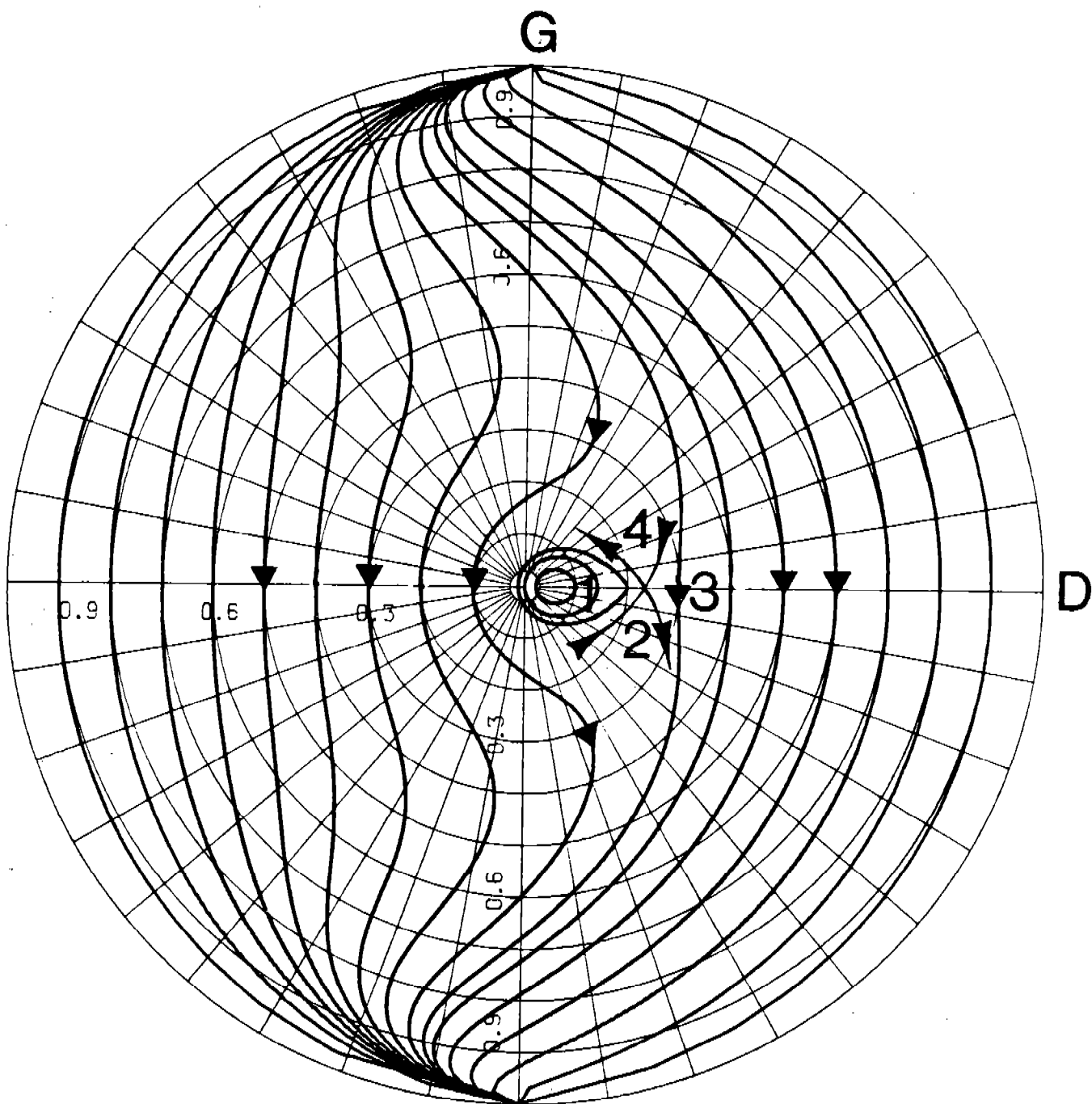


Figure 4 The phase plane for $\sigma=12.5 > \sigma_{cr} \approx 11.42$.

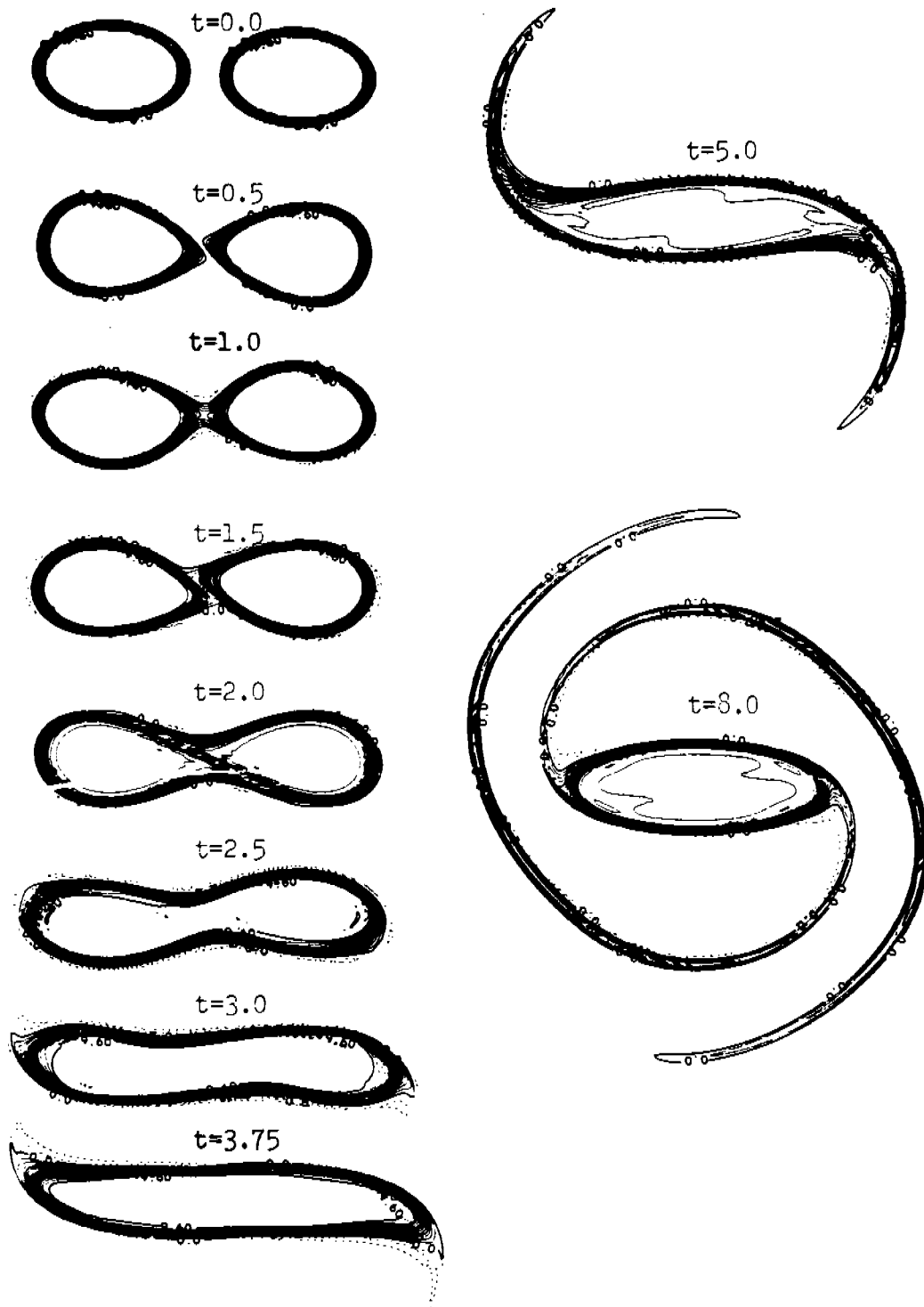


Figure 5 (legend on the following page)

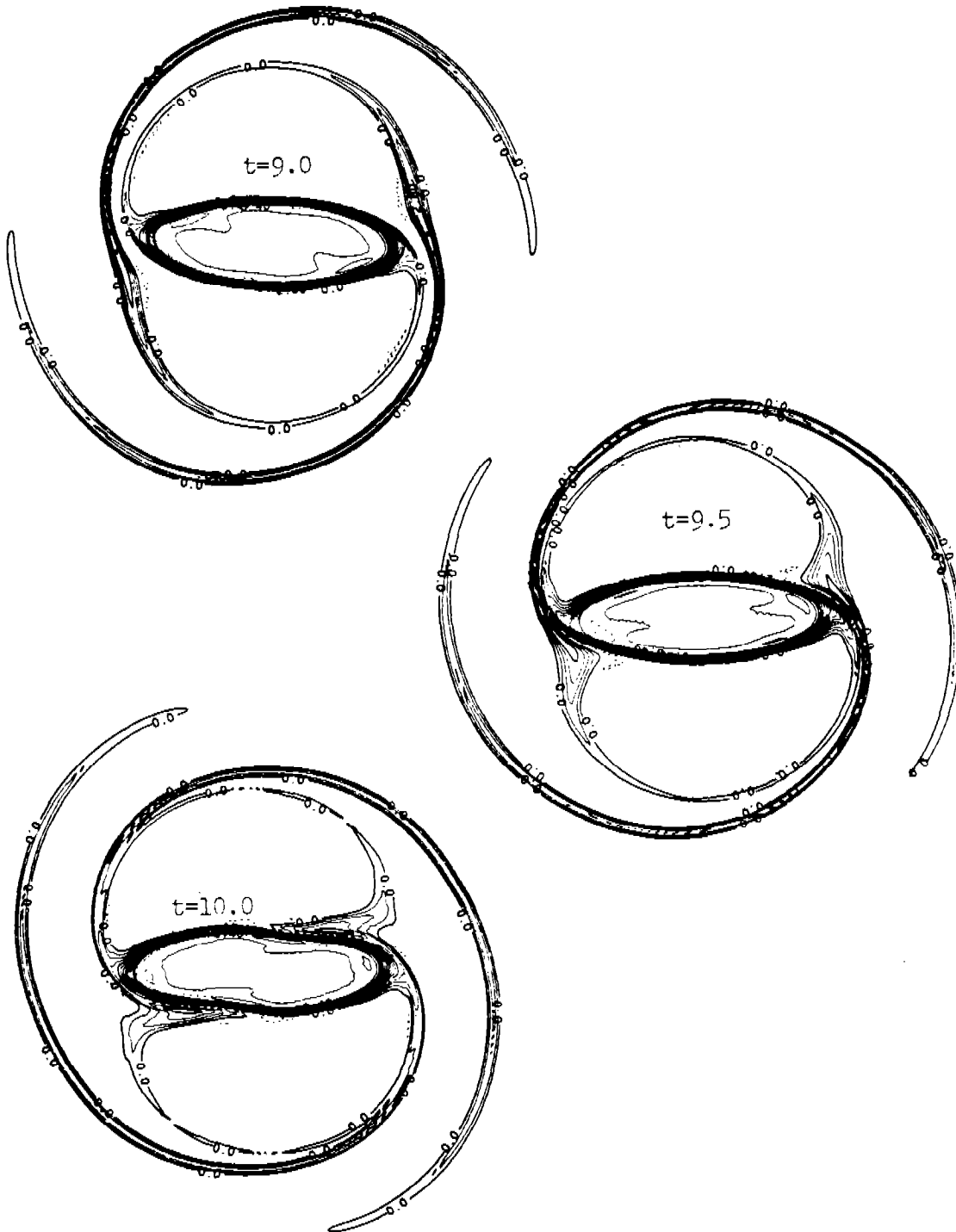


Figure 5 Equivorticity lines for the evolution of two $V(0.2, 10., 0.6255, 0.4)$ vortices on a 256^2 -mesh with $v_4 = 3.125 \times 10^{-8}$. The initial center-to-center distance is 1.44 and the box size is 2π . We have subtracted the bulk rotation in the pictures since it is irrelevant for our discussion. The contour interval is 0.6.

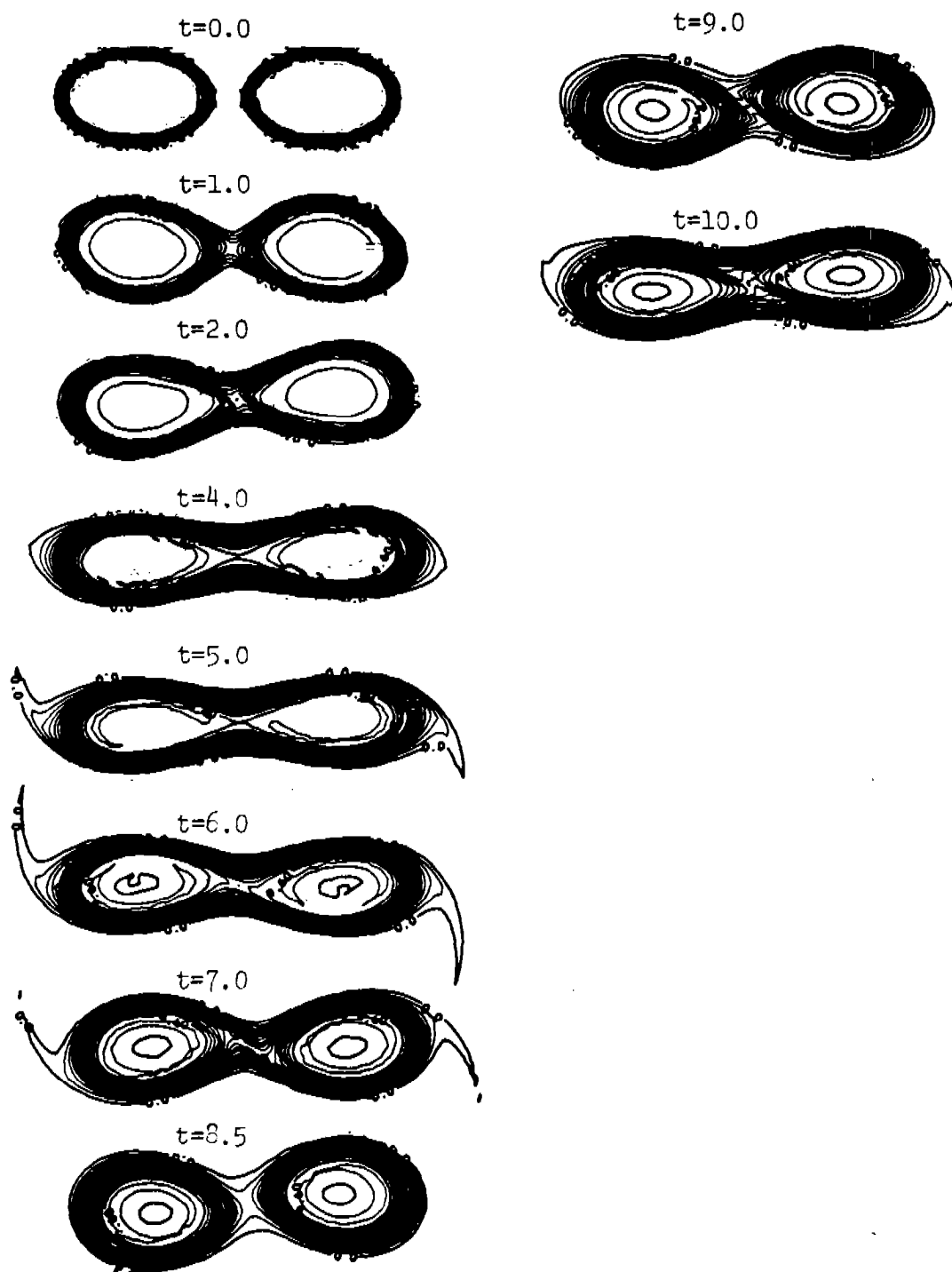


Figure 6 shows the evolution of the same two vortices as in Figure 5 but with the centroid separation 1.45. The calculation is performed on a 128^2 -mesh and the dissipation is 16 times stronger as dictated by the meshsize. Therefore the gradients quickly become more smooth than in Figure 5. The coarser mesh is justified by computational considerations. In order to observe near recurrence to the unstable steady state the initial centroid separation must be slightly smaller than the critical merger separation. This distance can only be found to high accuracy through bisection - a wasteful procedure. We found this distance on the 128^2 -mesh to three significant figures.

Stable Summation Methods for Elliptic Eigenfunction Expansions*

Harvey Diamond
West Virginia University

Mark Kon*
Boston University

Louise Raphael**
Howard University

Abstract: We develop stable methods for correct summation of expansions of data perturbed by error, using eigenfunctions associated with elliptic operators. The technique parallels that of Tikhonov for Sturm-Liouville operators in one dimension, using summation techniques which scale the summation parameter with error. Our analytic methods recover from the perturbed expansion a good approximation to the correct data, where the data are sufficiently regular.

Introduction: Some problems of engineering and mathematical physics can be formulated as that of solving the equation

$$Au = f, \quad f \in F$$

where A is an operator mapping a metric space U into a metric space F . These problems are called well-posed in the sense of Hadamard, if

- (a) there exists at least one solution u ;
- (b) there exists at most one solution u ;
- (c) there is continuous dependence of solutions $\{u\}$ upon the data $\{f\}$.

This last condition is called stability. A classical example of a problem which does not satisfy (c) is the Cauchy problem for the Laplace equation. Another is the heat-flow equation for negative time, namely

$$a^2 u_{xx} = u_t$$
$$u(x, t) = f(x), \quad t < T$$

This problem is unstable under small changes in the data f , in that small mean square errors in f produce large errors in u , both pointwise and in L^p .

One method of resolving such ill-posedness is the introduction of additional *a priori* assumptions on the solution u . The Russian mathematician A.N. Tikhonov [Til], in a fundamental paper, applied this idea to solution of ill-posed inverse problems. It was

* Research partially supported by the National Science Foundation

** Research supported by ARO Grant DAAG-84-G-0004

successfully used on the Cauchy problem for elliptic operators in pioneering work by Fritz John [Jo1], [Jo2].

This paper presents a regularization method for expansions in eigenfunctions of elliptic operators on \mathbf{R}^n . The following example illustrates a situation to which our results can be applied. If A is an elliptic operator in the variable $x \in \mathbf{R}^n$, the partial differential equation

$$\frac{\partial^k}{\partial t^k} \psi(x, t) = A\psi(x, t) \tag{1a}$$

on $(x, t) \in \mathbf{R}^n \times \mathbf{R}^+$ can be viewed as a dynamical system. Its normal modes are multiples of the eigenfunctions $\{u_\ell(x)\}$ of A and the solution can be written as

$$\psi(x, t) = \sum_{\ell} a_{\ell}(t) u_{\ell}(x), \tag{1b}$$

where time evolution of the dynamical variables $a_{\ell}(t)$ is governed by

$$\frac{d^k}{dt^k} a_{\ell}(t) = \lambda_{\ell} a_{\ell}(t),$$

with λ_{ℓ} the eigenvalue corresponding to $u_{\ell}(x)$. The above summation schematically describes a discrete or continuous spectral expansion. Equations such as (1) model such system as the flow of heat, and the time evolution of relativistic quantum mechanical and electromagnetic fields. Given the relationship between the “normal mode” dynamics and the “pointwise” dynamics expressed in (1), a natural question is how perturbations in configuration space (in which $\vec{a}(t) = (a_1, a_2, \dots)$ lies) correspond to changes in $\psi(x, t)$, and, more specifically, how such changes in $\psi(x, t)$ can be stably computed. This is clearly a question about pointwise behavior of perturbed expansions in eigenfunctions of elliptic operators.

Mathematically, our regularization method is a family of linear operators $\{\phi_{\epsilon}\}_{\epsilon>0}$ on $L^p(\mathbf{R}^n)$ along with a scaling function $\epsilon(\gamma), \gamma > 0$, such that if $\{f_{\gamma}\}, \gamma > 0$ is a net of $L^p(\mathbf{R}^n)$ perturbations of f with $\|f_{\gamma} - f\|_p < \gamma$, then $\phi_{\epsilon}(f_{\gamma})$ approaches f in L^p and pointwise when $\epsilon = \epsilon(\gamma)$ and $\gamma \rightarrow 0$. The results we obtain are applicable to the class of analytic multipliers $\phi_{\epsilon}(f) = \phi(\epsilon A)f$, where A is an elliptic operator on \mathbf{R}^n . Such regularization methods for expansions in eigenfunctions of ordinary differential operators have been widely studied. Results for partial differential operators are fewer. Our results apply to a large class of elliptic operators on \mathbf{R}^n , and can be used for both discrete and continuous spectra. We use extensively the analytic operator calculus (see [DS]), and the results and approach of [GK1] and [KR].

II. Previous Results and Problem Formulation.

Let $\alpha = (\alpha_1, \dots, \alpha_n)$ be a multiindex, and $D^{\alpha} \equiv (i)^{-|\alpha|} \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \dots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}$, $|\alpha| \equiv \alpha_1 + \dots + \alpha_n$. Consider the differential operator

$$A = \sum_{|\alpha| \leq m} b_{\alpha}(x) D^{\alpha} \equiv A_0 + B \tag{2}$$

on \mathbf{R}^n , where A_0 contains the leading terms and B is the remainder. We assume A_0 is constant coefficient positive elliptic, while the coefficients of B can be expressed as sums of functions in certain L^p spaces: $b_\alpha(x) \in L^{r_\alpha} + L^\infty$ ($|\alpha| < m$) where $d \equiv \sup_{|\alpha| < m} \{ \frac{n}{r_\alpha} + |\alpha| \} < m$. We choose for the domain of A the L^p -Sobolev space $\mathcal{L}_m^p = (1 - \Delta)^{-\frac{m}{2}} L^p$, $1 \leq p \leq \min r_\alpha$; if p is outside this range, then the domain must be smaller and A may not be densely defined. The resolvent $(z - A)^{-1}$ of A , if it exists, may be expressed as an integral operator with kernel denoted by $L_z(x, y)$, $x, y \in \mathbf{R}^n$. Sharp local bounds on the behavior of L_z were obtained in [GK2]. Specifically, let

$$h_{s,t}(x) = \begin{cases} |x|^{-s} & \text{if } s > 0, \\ -\ln|x| & \text{if } s = 0; \end{cases} \quad \begin{matrix} |x| \leq 1 \\ |x| \geq 1 \end{matrix} \quad (3)$$

The next theorem says that the kernel of the resolvent operator is bounded by an L^1 radially decreasing convolution kernel.

Theorem 1 (Gurarie, Kon): *Let $1 \leq p \leq \min r_\alpha$. There is a constant $C > 0$ such that the L^p -spectrum of A is contained in the parabolic domain $\Omega = \{z = \rho e^{i\theta} : C\rho^{\frac{d}{m}-1} \geq |\sin \frac{\theta}{2}|^{n+2}\}$ containing \mathbf{R}^+ . For $z \notin \Omega$ the kernel of $R_z = (z - A)^{-1}$ is estimated by*

$$|L_z(x, y)| \leq F(\rho, \theta) \rho^{\frac{n}{m}-1} h_{s,t}(\rho^{\frac{1}{m}}(x - y)), \quad (4)$$

where $t > n$, $s = \max(0, n - m)$ and

$$F(z) = F(\rho, \theta) = \frac{C}{|\sin \frac{\theta}{2}|^{n+2}} \left(1 - \frac{C\rho^{\frac{d}{m}-1}}{|\sin \frac{\theta}{2}|^{n+2}} \right)^{-1}. \quad (5)$$

Furthermore, the kernel $L_z^{(1)}$ of $R_z - R_z^0$, where $R_z^0 = (z - A_0)^{-1}$, satisfies

$$L_z^{(1)}(x, y) \leq \frac{C}{|\sin \frac{\theta}{2}|^{n+2}} F(\rho, \theta) \rho^{\frac{d+n}{m}-2} h_{s',t}(\rho^{\frac{1}{m}}(x - y)),$$

with $s' = \max(n - 2m + d, 0)$.

We now define our class of regularization operators, which are expressed in terms of an analytic multiplier $\phi(z)$ with $\phi(0) = 1$. Let

$$B_\gamma = \{z : |\arg z| \leq \gamma\}, \quad D_r = \{z : |z| \leq r\} \quad (6)$$

By Theorem 1, the spectrum $\sigma(A)$ is contained in $B_\gamma \cup D_r$ for r sufficiently large. Let $\phi(z)$ be analytic in a domain $\mathcal{D} \subset \mathbf{C}$ containing $B_\gamma \cup D_r$, and Γ be the positively oriented contour consisting of the boundary $\partial(D_r \cup B_\gamma)$. Using the analytic operator calculus we can define the operator $\phi(\epsilon A) : L^p(\mathbf{R}^n) \rightarrow L^p(\mathbf{R}^n)$ by

$$\phi(\epsilon A)f(x) = \frac{1}{2\pi i} \int_\Gamma \phi(z)(z - \epsilon A)^{-1} f(x) dz,$$

where ϵ is small enough that the spectrum of ϵA lies within Γ .

To guarantee $\phi(\epsilon A)$ is well defined we suppose

$$(a) \quad \int_{R_{\gamma_1}} \left| \frac{\phi(z)}{z} \right| |dz| < \infty \text{ for any ray } R_{\gamma_1} = \{z: \arg z = \gamma_1\}, \quad |\gamma_1| \leq \gamma \quad (7a)$$

$$(b) \quad \int_{G_r} \left| \frac{\phi(z)}{z} \right| |dz| \xrightarrow{r \rightarrow \infty} 0 \text{ where } G_r = \{z: |z| = r, |\arg z| \leq \gamma\} \quad (7b)$$

For our regularization methods, we must impose the following stronger integrability condition:

$$\int_{\Gamma} |\phi(z)| |z|^{-\delta} |dz| < \infty \quad (8)$$

for all $\delta > 0$.

It is interesting to note the connection of the regularization operators $\phi(\epsilon A)$ with the application of multipliers for summing expansions. Formally,

$$\phi(\epsilon A) \sum_{\lambda} a_{\lambda} u_{\lambda} = \sum_{\lambda} \phi(\epsilon \lambda) a_{\lambda} u_{\lambda},$$

with u_{λ} a generalized eigenfunction of A and λ its associated eigenvalue. The operator formulation on the left side is more general, in that it can be applied independently of detailed spectral considerations.

III. Main Result and an Application

We can now formulate our problem. Let $f \in L^p(\mathbf{R}^n)$, $1 \leq p < \infty$, and $\{f_{\gamma}\}$, $\gamma > 0$, denote a net of functions in $L^p(\mathbf{R}^n)$ with $\|f - f_{\gamma}\|_p < \gamma$. We wish to determine a sharp scaling of ϵ and γ which guarantees that as $\gamma \rightarrow 0$, $\phi(\epsilon A)f_{\gamma} \rightarrow f$ in $L^p(\mathbf{R}^n)$, $1 \leq p < \infty$, and pointwise on the Lebesgue set of f .

The following theorem provides a large class of L^p -regularizing operators of the form $\phi(\epsilon A)$ and a sharp condition on the associated scaling functions.

Theorem 2: *Let A be an elliptic operator of order m on \mathbf{R}^n satisfying the conditions above, and $\phi(z)$ be analytic on the spectrum of A , satisfying the boundedness conditions (7) and (8). Let $p > \frac{n}{m}$, $f \in L^p(\mathbf{R}^n)$, and $\{f_{\gamma}\}$, $\gamma > 0$, be a net of $L^p(\mathbf{R}^n)$ functions, with $\|f - f_{\gamma}\|_p < \gamma$. If*

$$\gamma \epsilon^{-\frac{n}{mp}} \xrightarrow{\gamma \rightarrow 0} 0 \quad (9)$$

then $\phi(\epsilon A)f_{\gamma} \rightarrow f$ in $L^p(\mathbf{R}^n)$ ($1 \leq p < \infty$) and pointwise on the Lebesgue set of f . Furthermore, this scaling of ϵ with γ is sharp in that if (9) fails, then for each A there are $f, \{f_{\gamma}\}$ such that $\phi(\epsilon A)f_{\gamma} \not\rightarrow f$ on the Lebesgue set of f .

The proof uses the fact that for f in L^p , $1 \leq p < \infty$, $\phi(\epsilon A)f(x)$ converges as $\epsilon \rightarrow 0$ to $f(x)$ in L^p norm ($1 \leq p < \infty$) and pointwise on the Lebesgue set of f ($1 \leq p \leq \infty$); it also depends on Theorem 1, the Minkowski integral inequality, and Young's inequality. The proof is omitted for brevity, and will appear elsewhere.

As an application, consider a function $f(\vec{t})$ in a multiparameter time domain ($\vec{t} = (t_1, \dots, t_n)$),

$$f(\vec{t}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} \tilde{f}(\vec{\omega}) e^{i\vec{t} \cdot \vec{\omega}} d\vec{\omega}. \quad (10)$$

How do frequency domain errors (in $\tilde{f}(\vec{\omega})$) propagate into time domain ($f(\vec{t})$) errors? In general, small function space or pointwise errors in \tilde{f} will lead to unbounded errors in pointwise estimation of f .

This type of problem has applications in many types of situations; one example is the problem of waves radiating from an antenna with fixed frequency ω , and a given aperture illumination $f(x)$. Small perturbations in $f(x)$ can lead to unbounded errors in the local radiation field intensity.

The following provides a regularization procedure for ameliorating the large pointwise errors in $f(t)$ above; we choose to work with L^2 for simplicity.

Corollary 2.1. Let \tilde{f}_γ be an L^2 -perturbation of \tilde{f} , of size γ , i.e., $\|\tilde{f} - \tilde{f}_\gamma\|_2 \leq \gamma$

$$f_{\epsilon, \gamma}(\vec{t}) \equiv \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} \tilde{f}_\gamma(\vec{\omega}) \phi(\epsilon \vec{\omega}^2) e^{i\vec{\omega} \cdot \vec{t}} d\vec{\omega}, \quad (11)$$

where ϕ is a function analytic in a region $B_\gamma \cup D_r$ ($\gamma, r > 0$; see eq. (6)), and satisfies (7) and (8). Let $f(t)$ be the inverse Fourier transform of \tilde{f} . If the scaling

$$\gamma \epsilon^{-\frac{n}{4}} \xrightarrow{\gamma \rightarrow 0} 0 \quad (12)$$

holds, then the pointwise error in $f_{\epsilon, \gamma}(\vec{t})$ vanishes as $\epsilon \rightarrow 0$.

That is, if the summation parameter ϵ is scaled correctly with L^2 -error γ , pointwise error in the time domain vanishes with function space error in the frequency domain. We remark that the exponent $-\frac{n}{4}$ in (12) arises from the fact that the expansion in (10) is in fact in eigenfunctions of the Laplacian. That is,

$$f_{\epsilon, \gamma} = \phi(-\epsilon \Delta) f_\gamma.$$

Bibliography

- [AT] Arsenin, V.Y. and A.N. Tikhonov, *Solutions of ill-posed problems*, Winston, 1977.
- [DS] Dunford, N. and J.T. Schwartz, *Linear Operators*, Interscience, New York, 1958.
- [DKR] Diamond, H., M. Kon and L. Raphael, Stable summation methods for a class of singular Sturm-Liouville expansions, *Proc. Amer. Math. Soc.* (2) 81 (1981), 279-286.

- [GK1] Gurarie, D. and M. Kon, Radial bounds for perturbations of elliptic operators, *J. Functional Analysis* **56** (1984), 99-123.
- [GK2] Gurarie, D. and M. Kon, Resolvents and regularity properties elliptic operators. *Operator Theory: Advances and Applications*, 151-162, Birkhäuser Verlag, Basel, 1983.
- [Jo1] John, F., A note on "improper" problems in partial differential equations, *Comm. Pure and Appl. Math.*, **8** (1955).
- [Jo2] John, F., Continuous dependence on data for solutions with a prescribed bound, *Comm. Pure and Appl. Math.*, (**4**) **13** (1960).
- [Ko] Kon, M., Regularity properties of Schrödinger operators on a domain of \mathbf{R}^n , *Differential Equations*, I.W. Knowles and R.T. Lewis, Eds., Elsevier Science Publishers, 1984.
- [KR] Kon, M. and L. Raphael, New multiplier methods for summing classical eigenfunction expansions, *J. Differential Equations*, **50** (1983), 391-406.
- [Ti1] Tikhonov, A.N., The stability of inverse problems, *Doklady Akad. Nauk SSSR* (**5**) **39** (1943).
- [Ti2] Tikhonov, A.N., Stable methods for the summation of Fourier series, *Soviet Math. Dokl.* **5** (1964), 641-644.

ON THE USE OF PIECEWISE-POLYNOMIALS FOR THE APPROXIMATION OF CAUCHY SINGULAR INTEGRALS

Apostolos Gerasoulis

Department of Computer Science

Rutgers University

New Brunswick, NJ 08903

ABSTRACT: *In this paper we propose piecewise-polynomial methods for the approximation of Cauchy principal value integrals and develop a simple, efficient and numerically stable algorithm for the evaluation of the weights of the resulting piecewise-polynomial quadratures. We present two examples to illustrate the advantages of these quadratures versus the Gauss-Jacobi quadratures.*

1. INTRODUCTION

In this paper we consider the numerical approximation of the Cauchy principal value integral

$$I(g; s) = \int_{-1}^1 w(t) \frac{g(t)}{t-s} dt = \lim_{\epsilon \rightarrow 0} \left\{ \int_{-1}^{s-\epsilon} + \int_{s+\epsilon}^1 \right\} w(t) \frac{g(t)}{t-s} dt \quad (1)$$

where $g(t)$ is a Holder continuous function in $[-1, 1]$ and $w(t)$ is a positive integrable weight function given by

$$w(t) = (1-t)^{-\alpha}(1+t)^{-\beta} \quad (2)$$

where α and β are constants and $\alpha, \beta < 1$. There has been an increasing interest in the numerical approximation of (1), particularly since such approximations may be used to solve the Cauchy Singular Integral Equation (CSIE):

$$a(s)w(s)g(s) + \frac{b(s)}{\pi} I(g; s) = f(s), \quad |s| < 1 \quad (3)$$

where $a(s)$, $b(s)$ and $f(s)$ are given input functions and $g(t)$ is the unknown function. CSIE's arise in areas such as aerodynamics, fluid and fracture mechanics, wave-guide theory, scattering and other areas (e.g. Erdogan et. al. [6], Gerasoulis [8], Lotz [12], Miller and Keer [13]).

The majority of numerical methods proposed for (1) are *global* methods, usually based on

orthogonal polynomial approximations (e.g. Paget and Elliott [14], Erdogan et. al. [6], Welstead [17]). Very little attention has been paid to *local* methods, mainly piecewise-polynomial approximations, even though such methods were used as early as 1950 for the special case $w(t) = 1$ (e.g. Lotz [12], Ivanov [10]).

It is well known that global methods converge very fast for differentiable input functions with "small" derivatives. However, there are two difficulties associated with their implementation. First, although orthogonal polynomials exist for the approximation of (1) (Elliott [5]), similar polynomials may not exist for (3), since nonclassical weight functions might arise for certain $b(s)$ (Welstead [17, p. 115]). Only for special cases of $b(s)$ the existence of orthogonal polynomials has been proven (Elliott [5]). Even if such polynomials exist for (3), considerable computational effort is needed to generate their recursion coefficients (e.g. Welstead [17, p. 115]). Second, since the node points are generally chosen as the zeros of an orthogonal polynomial, global methods are not appropriate for integrals with input functions that behave "badly" in some subinterval $[a, b]$ of $[-1, 1]$ (e.g. Paget and Elliott [14, pp. 381-384]). For such integrals, a numerical method with no restriction on the choice of node points would have to be used in order to concentrate the nodes in $[a, b]$. This is not possible with global methods.

In contrast, local methods based on piecewise-polynomial quadratures afford a flexible choice of the node points. However, their implementation suffers from the lack of efficient numerical algorithms for the computation of the quadrature weights. In this paper we address this issue and propose a simple, efficient and numerically *stable* method for the computation of the quadrature weights.

In section 2, we develop a quadrature for $I(g; s)$, by approximating $g(t)$ with piecewise-linear polynomials. We also develop an algorithm for the evaluation of the quadrature weights by splitting the interval $[-1, 1]$ into two subintervals $[-1, 0]$ and $[0, 1]$, and in each subinterval, we expand the nonsingular parts $(1-t)^{-\alpha}$ and $(1+t)^{-\beta}$ into Taylor's series and integrate the singular parts of $w(t)$ "exactly". The quadrature weights are computed via two fast-converging series. Finally in section 3, two numerical examples are presented. In the first example we extend the piecewise-linear quadrature to weakly singular integrals of the form $\int_{-1}^1 w(t)g(t)dt$, where $w(t)$ is defined in (2). We illustrate the advantages of these quadratures versus the Gauss-Jacobi quadratures by choosing $g(t) = \sqrt{|t|}$ and also describe how the analysis for piecewise-linear quadrature may be extended to higher order approximations (e.g. piecewise-quadratic, etc.). In the second example we demonstrate the usefulness of the piecewise-linear quadrature for the approximation of integrals with weight functions of the form with $w(t) = (1-t)^{-\alpha}(1+t)^{-\beta}\Omega(t)$, where $\Omega(t)$ is a positive continuous function. Such weight functions arise in the solution of CSIE's with variable coefficients (e.g. Elliott [5], Welstead [17]).

2. PIECEWISE-POLYNOMIAL QUADRATURES

We begin this section with some preliminary mathematics. We rewrite (1) as

$$I(g; s) = J(g; s) + g(s)q_0(s), \quad J(g; s) = \int_{-1}^1 w(t) \frac{g(t) - g(s)}{t - s} dt, \quad q_0(s) = \int_{-1}^1 \frac{w(t)}{t - s} dt \quad (4)$$

and assume that $q_0(s)$ exists. A sufficient condition for its existence is that $w(t)$ is Holder continuous in every open subinterval of $[-1, 1]$, which is true for the weight function in (2). The function $q_0(s)$ satisfies

$$q_0(s) = -\pi \cot(\pi\alpha) w(s) - 2^{-(\alpha+\beta)} \frac{\Gamma(-\alpha)\Gamma(-\beta+1)}{\Gamma(-\alpha-\beta+1)} F(1, -(\alpha+\beta); 1+\alpha; \frac{1-s}{2}), \quad \alpha \neq 0, -1, \dots \quad (5)$$

where $F(1, -(\alpha+\beta); 1+\alpha; \frac{1-s}{2})$ is the hypergeometric and $\Gamma(z)$ the gamma function (e.g. Tricomi [16]). For several important special cases the hypergeometric function may be simplified further. For example, if $\alpha+\beta = \kappa$ and κ is an integer, then

$$2^{-(\alpha+\beta)} \frac{\Gamma(-\alpha)\Gamma(-\beta+1)}{\Gamma(-\alpha-\beta+1)} F(1, -(\alpha+\beta); 1+\alpha; \frac{1-s}{2}) = -2^{-\kappa} \frac{\pi}{\sin(\pi\alpha)} P_{\kappa}^{(\alpha, \beta)}(s) \quad (6)$$

where $P_{\kappa}^{(\alpha, \beta)}(s)$ is the Gauss-Jacobi polynomial of degree κ . We are now ready to introduce the piecewise-linear quadrature.

2.1. Piecewise-Linear Approximations

We subdivide $[-1, 1]$ into n subintervals $[t_i, t_{i+1}]$, $t_i < t_{i+1}$, $i = 0(1)n-1$, $t_0 = -1$, $t_n = 1$, and define the stepsize $h_i = t_{i+1} - t_i$, $i = 0(1)n-1$. The piecewise-linear polynomial approximation $g_n(t)$ of $g(t)$ is defined by

$$g_n(t) = a_i(t - t_i) + b_i, \quad \text{if } t \in [t_i, t_{i+1}], \quad a_i = \frac{g(t_{i+1}) - g(t_i)}{h_i}, \quad b_i = g(t_i), \quad i = 0(1)n-1. \quad (7)$$

Furthermore, we set

$$I_i = \int_{t_i}^{t_{i+1}} w(t) dt, \quad I_i(s) = \int_{t_j}^{t_{j+1}} \frac{w(t)}{t-s} dt, \quad A_i(s) = a_i(s - t_i) - a_j(s - t_j) + b_i - b_j, \quad (8)$$

for $i = 0(1)n-1$, and present the following theorem.

Theorem 1: (i) For $s \in [t_j, t_{j+1}]$, $j \in \{0, 1, \dots, n-1\}$, we have $A_j(s)I_j(s) = 0$ and

$$J(g_n; s) = \sum_{i=0}^{n-1} \{a_i I_i + A_i(s) I_i(s)\}. \quad (9)$$

(ii) If $g(t)$ is continuously differentiable, then $J(g_n; s)$ converges uniformly to $J(g; s)$.

Proof: (i) We first show that $A_j(s)I_j(s) = 0$ for $s \in [t_j, t_{j+1}]$. Since $w(t)$ is Holder continuous in every closed subinterval of $(-1, 1)$, then there exists an $\epsilon > 0$ so that the integrals

$$\int_{t_j}^{t_{j+1}} \frac{w(t)}{t-s} dt = \int_{t_j}^{t_{j+1}} \frac{w(t)-w(s)}{t-s} dt + w(s) \int_{t_j}^{t_{j+1}} \frac{dt}{t-s}, \quad j=1(1)n-2$$

$$\int_{t_j}^{t_{j+1}} \frac{w(t)}{t-s} dt = \int_{s-\epsilon}^{s+\epsilon} \frac{w(t)-w(s)}{t-s} dt + \int_{t_j}^{s-\epsilon} \frac{w(t)}{t-s} dt + \int_{s+\epsilon}^{t_{j+1}} \frac{w(t)}{t-s} dt, \quad j=0, n-1,$$

are bounded. Moreover, $A_j(s) = 0$, for $s \in (t_j, t_{j+1})$, $j = 0(1)n-1$ and therefore

$$A_j(s) \int_{t_j}^{t_{j+1}} \frac{w(t)}{t-s} dt = 0 \quad \text{for } s \in (t_j, t_{j+1}). \quad (10)$$

Similarly, for $s = t_{j+1}$, $A_j(s) = A_{j+1}(s) = 0$ and

$$A_j(s) \int_{t_j}^{t_{j+2}} \frac{w(t)}{t-t_{j+1}} dt = 0, \quad j = 0(1)n-2. \quad (11)$$

By substituting $g_n(t)$ in the second equation in (4), by rewriting (4) as

$$J(g_n; s) = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} w(t) \frac{a_i(t-t_i)+b_i - a_j(s-t_j)-b_j}{t-s} dt \quad (12)$$

and by using (10), (11) and (12) we derive (9).

(ii) The proof of convergence shown by Stewart [15] for $\alpha = \beta = 1/2$, also applies to the weight function (2) for all $\alpha, \beta < 1$. \square

We now consider the quadrature approximation for $I(g; s)$. We define

$$l_i(s) = \begin{cases} I_i(s) & \text{if } s \notin [t_i, t_{i+1}] \\ 0 & \text{otherwise} \end{cases}, \quad v_i = \frac{I_i + (s-t_i)l_i(s)}{h_i}, \quad i = 0(1)n-1, \quad (13)$$

$$w_0(s) = l_0(s) - v_0(s), \quad w_n(s) = v_{n-1}(s), \quad w_i(s) = v_{i-1}(s) + l_i(s) - v_i(s), \quad i = 1(1)n-1, \quad (14)$$

and present the following corollary, which can be shown directly from Theorem 1.

Corollary: The piecewise-linear quadrature for $I(g; s)$ is given by

$$I(g_n; s) = \sum_{i=0}^n w_i(s) g(t_i) + \{ q_0(s) - \sum_{i=0}^{n-1} l_i(s) \} g_n(s) \quad (15)$$

where $g_n(s)$, $q_0(s)$ and $w_i(s)$ are defined in (7), (4) and (14) respectively. \square

It is clear from Theorem 1 that the main difficulty with the piecewise-polynomial quadrature (9) is the evaluation of I_i and $I_i(s)$. Several authors have considered particular cases of $w(t)$. Flugge-Lotz [12], Atkinson [2] and others (see the references in Stewart [15] and Ivanov [10]) have studied the case $w(t) = 1$. For this case, we see that the integrals in (8) have a closed form expression, e.g. $I_i = t_{i+1} - t_i$ and $I_i(s) = \ln |(t_{i+1} - s)/(t_i - s)|$. Closed form expressions have also been derived for the cases $\alpha = \beta = \pm 1/2$ by Gerasoulis [7], [8], Gerasoulis and Srivastav [9], Jen and Srivastav [11]. Such closed form expressions do not exist in general and a numerical method must be used. However, since the integrands in (8) are singular at ± 1 , classical quadratures are not appropriate for their approximation. It is only recently that an attempt was made to develop numerical methods for the evaluation I_i and $I_i(s)$: Miller and Keer [13] suggested an interesting technique for estimating I_i and $I_i(s)$. They first expand the weight function $w(t) = (1-t)^{-\alpha}(1+t)^{-\beta}$ into a Taylor's series in $(-1, 1)$ and then integrate the series "exactly". However, since the Taylor's series is unbounded at ± 1 , the integrated series converges slowly near these points.

In the next section, we introduce a simple, efficient and numerical stable method for the evaluation of I_i and $I_i(s)$. The underlying idea in the evaluation of I_i and $I_i(s)$ is similar to "product integration" which has been successfully used in quadrature approximations of weakly singular integrals (Atkinson [1, p. 272]). Thus, instead of expanding $w(t) = (1-t)^{-\alpha}(1+t)^{-\beta}$ in $(-1, 1)$ as it has been proposed by Miller and Keer [13], we first split the interval $[-1, 1]$ into two subintervals $[-1, 0]$ and $[0, 1]$ and expand the part of $w(t)$ which consists of a C^∞ function and integrate the other part. Note that $(1-t)^{-\alpha}$ is C^∞ in $[-1, 0]$ and that $(1+t)^{-\beta}$ is C^∞ in $[0, 1]$. Clearly, this approach would be practical only if the integrated part can be computed very efficiently to within a given tolerance. Fortunately, as we shall see in the next two sections, this is possible for the weight function defined in (2).

2.2. The Numerical Evaluation of I_i

In Theorems 2 and 3 we present a series approximation for I_i and analyze its convergence. Note that each step in the series is computed recursively and therefore very few computations are performed per summation step.

Theorem 2: *The integrals I_i , $i = 0(1)n-1$, defined in (8) can be evaluated from (a), (b) and (c) below,*

(a) If i is such that $0 \leq t_i < t_{i+1}$, and $p \in [0, 1]$,

$$I_i = (1+p)^{-\beta} \sum_{k=0}^{\infty} (-1)^k \frac{\Gamma(k+\beta)}{\Gamma(\beta)k!(1+p)^k} I_{i,k} \quad (16)$$

$$I_{i,k} = \frac{k(1-p)}{1+k-\alpha} I_{i,k-1} - \frac{1}{1+k-\alpha} \{ (1-t_{i+1})^{1-\alpha} (t_{i+1}-p)^k - (1-t_i)^{1-\alpha} (t_i-p)^k \}, \quad I_{i,-1} = 0 \quad (17)$$

(b) If i is such that $t_i < t_{i+1} \leq 0$, and $p \in [-1, 0]$,

$$I_i = (1-p)^{-\alpha} \sum_{k=0}^{\infty} \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)k!(1-p)^k} I_{i,k} \quad (18)$$

$$I_{i,k} = -\frac{k(1+p)}{1+k-\beta} I_{i,k-1} + \frac{1}{1+k-\beta} \{ (1+t_{i+1})^{1-\beta} (t_{i+1}-p)^k - (1+t_i)^{1-\beta} (t_i-p)^k \}, \quad I_{i,-1} = 0. \quad (19)$$

(c) If i is such that $t_i \leq 0 \leq t_{i+1}$, then this interval is split into two subintervals $[t_i, 0]$ and $[0, t_{i+1}]$ and I_i is evaluated by using (a) and (b) in each subinterval.

Proof : We prove only part (a) since the proof for part (b) and (c) is similar. The Taylor's series expansion of $(1+t)^{-\beta}$ is given by

$$(1+t)^{-\beta} = (1+p)^{-\beta} \sum_{k=0}^{\infty} (-1)^k \frac{\Gamma(k+\beta)}{\Gamma(\beta)k!(1+p)^k} (t-p)^k \quad (20)$$

which converges for all points $p \in [0, 1]$. By multiplying with $(1-t)^{-\alpha}$ and by integrating the last equation, we derive (16), where

$$I_{i,k} = \int_{t_i}^{t_{i+1}} \frac{(t-p)^k}{(1-t)^\alpha} dt. \quad (21)$$

To show (17), we consider the indefinite integral $I_{\bullet,k}$ of $I_{i,k}$ and integrate it by parts

$$I_{\bullet,k} = \int \frac{(t-p)^k}{(1-t)^\alpha} dt = -\frac{1}{1-\alpha} \{ (t-p)^k (1-t)^{1-\alpha} - k \int \frac{[(1-p)-(t-p)](t-p)^{k-1}}{(1-t)^\alpha} dt \} \quad (22)$$

which reduces to the following remarkable recursion

$$I_{\bullet,k} = \frac{k(1-p)}{1+k-\alpha} I_{\bullet,k-1} - \frac{(1-t)^{1-\alpha} (t-p)^k}{1+k-\alpha}. \quad (23)$$

Then, (17) is derived by taking the integration limits in (23). Furthermore, since $\alpha, \beta < 1$, then

$$0 \leq \frac{k(1-p)}{1+k-\alpha} \leq 1, \text{ if } p \in [0, 1] \quad \text{and} \quad 0 \leq \frac{k(1+p)}{1+k-\beta} \leq 1, \text{ if } p \in [-1, 0],$$

which imply that the recursions (17) and (19) are numerically stable [3, p. 17]. \square

We now consider the rate of convergence of the partial sums $I_i^{(N)}$, where

$$I_i^{(N)} = \sum_{k=0}^N d_k, \quad d_k = (-1)^k \frac{\Gamma(k+\beta)}{\Gamma(\beta)k!(1+p)^{k+\beta}} I_{i,k}, \quad R_N = I_i - I_i^{(N)} = \sum_{k=N+1}^{\infty} d_k. \quad (24)$$

Theorem 3: If $p = t_i$ or $p = t_{i+1}$, then the remainder R_N of the series in (16) satisfies:

$$|R_N| \leq \frac{\lambda |d_N|}{1-\lambda} \leq \frac{\lambda^{N+1-M} |d_M|}{1-\lambda}, \quad \text{for all } N \geq M = \left\lceil \frac{1-\beta}{2} \right\rceil. \quad (25)$$

where $\lambda = \frac{h_i}{1+p} < 1$.

Proof : We will show that $|d_k| \leq \lambda |d_{k-1}|$ and $\lambda < 1$ for all $k \geq M$. Since $\beta = 0, -1, \dots$, implies $d_k = 0$, for all $k > -\beta$, we need to consider only the case $\beta \neq 0, -1, \dots$. We see that the assumption $p = t_{i+1}$ or $p = t_i$ and (21) imply

$$|I_{i,k}| \leq (t_{i+1} - t_i) |I_{i,k-1}|, \quad \text{and from (24)} \quad |d_k| \leq \frac{|k-1+\beta|h_i}{k(1+p)} |d_{k-1}|. \quad (26)$$

Furthermore, since $\beta < 1$ and $0 < p \leq 1$ then $|k-1+\beta| < k$ for all $k \geq M = \left\lceil \frac{1-\beta}{2} \right\rceil$, and by using (26), $|d_k| \leq \lambda |d_{k-1}|$, where $\lambda = \frac{h_i}{1+p} < 1$. From the last equation in (24)

$$|R_N| \leq \sum_{k=N+1}^{\infty} |d_k| \leq \sum_{k=N+1}^{\infty} \lambda^{k-N-1} |d_{N+1}| = \frac{|d_{N+1}|}{1-\lambda} \leq \frac{\lambda |d_N|}{1-\lambda}, \quad \text{for } N \geq M,$$

where we have used the sum of the the geometric series [3, p. 63]. \square

This Theorem implies that the partial sums $I_i^{(N)}$ converge very fast to I_i . For example, if $\beta = 1/2$ and $h_i = .2$, i.e. $n = 10$ subintervals, then $|R_N| \leq 4 \times 10^{-9}$ for $N \leq 9$, while if $h_i = .02$, i.e. $n = 100$ subintervals, then $|R_N| \leq 3 \times 10^{-9}$ for $N \leq 3$. The bound for the remainder of the series in (18) is similar to (25).

2.3. The Numerical Evaluation of $I_i(s)$

The piecewise-linear quadrature derived in Theorem 1, also requires the evaluation of $I_i(s)$, $i = 0(1)n-1$, for $s \notin [t_i, t_{i+1}]$. In the next two Theorems, we present a series approximation for $I_i(s)$ and analyze its convergence. Note again that each step in the series is computed recursively and therefore very few computations are performed per summation step (i.e. 9 multiplications and 5 additions per step).

Theorem 4: The integrals $I_i(s)$, $i = 0(1)n-1$ defined in (8), can be evaluated for $s \notin [t_i, t_{i+1}]$ from (a), (b) and (c) below, where the point p is chosen as

$$p = \begin{cases} t_i & \text{if } s > t_{i+1} \\ t_{i+1} & \text{if } s < t_i \end{cases}$$

(a) If i is such that $0 \leq t_i < t_{i+1}$, then

$$I_i(s) = \sum_{k=0}^{\infty} B_k(s) I_{i,k}(s), \quad I_{i,k}(s) = \frac{I_{i,k}}{(s-p)^k} \quad (27)$$

where $I_{i,k}$ is defined in (17) and $B_k(s)$ has a closed form expression for $\beta = 0, -1, -2, \dots$, e.g. if $\beta = 0$ then $B_k(s) = 1/(p-s)$, while for $\beta \neq 0, -1, -2, \dots$,

$$B_k(s) = B_{k-1}(s) + b_k(s), \quad b_k(s) = -\frac{(k-1+\beta)}{k} \left(\frac{s-p}{1+p} \right) b_{k-1}(s), \quad b_0(s) = \frac{(1+p)^{-\beta}}{p-s}, \quad (28)$$

$$B_{-1}(s) = 0.$$

If numerical underflow occurs in the evaluation of $I_{i,k}(s)$ by (27) then the following equivalent recursion should be used:

If $\frac{|1-p|}{|s-p|} \leq 1$, then set $I_{i,-1}(s) = 0$ and evaluate

$$I_{i,k}(s) = \frac{k}{1+k-\alpha} \left(\frac{1-p}{s-p} \right) I_{i,k-1}(s) - \frac{1}{1+k-\alpha} \left\{ (1-t_{i+1})^{1-\alpha} \left(\frac{t_{i+1}-p}{s-p} \right)^k - (1-t_i)^{1-\alpha} \left(\frac{t_i-p}{s-p} \right)^k \right\}, \quad (29)$$

for $k = 0, 1, \dots$. If $\frac{|1-p|}{|s-p|} > 1$ then choose a sufficiently large N , set $I_{i,N}(s) = 0$ and evaluate $I_{i,k}(s)$, $k = (N-1)(-1)1$ by using (29) backwards.

(b) If i is such that $t_i < t_{i+1} \leq 0$, then $I_i(s)$ is evaluated from (27), where $I_{i,k}$ is defined in (19) and $B_k(s)$ has a closed form expression for $\alpha = 0, -1, -2, \dots$, while for $\alpha \neq 0, -1, -2, \dots$,

$$B_k(s) = B_{k-1}(s) + b_k(s), \quad b_k(s) = \frac{(k-1+\alpha)}{k} \left(\frac{s-p}{1-p} \right) b_{k-1}(s), \quad b_0(s) = \frac{(1-p)^{-\alpha}}{p-s}, \quad (30)$$

$B_{-1}(s) = 0$. Similarly, if $\frac{|1+p|}{|s-p|} \leq 1$, then set $I_{i,-1}(s) = 0$ and evaluate

$$I_{i,k}(s) = -\frac{k}{1+k-\beta} \left(\frac{1+p}{s-p} \right) I_{i,k-1}(s) + \frac{1}{1+k-\beta} \left\{ (1+t_{i+1})^{1-\beta} \left(\frac{t_{i+1}-p}{s-p} \right)^k - (1+t_i)^{1-\beta} \left(\frac{t_i-p}{s-p} \right)^k \right\}, \quad (31)$$

for $k = 0, 1, \dots$, while if $\frac{|1+p|}{|s-p|} > 1$ then use (31) backwards.

(c) If i is such that $t_i \leq 0 \leq t_{i+1}$, then this interval is split into two subintervals $[t_i, 0]$ and $[0, t_{i+1}]$ and $I_i(s)$ is evaluated by using (a) and (b) in each subinterval.

Proof: We prove only part (a) since the proof for (b) and (c) is similar. From the definition of p and the fact that $s \notin [t_i, t_{i+1}]$, we have $|t-p| < |s-p|$ for $t_i \leq t \leq t_{i+1}$ and

$$\frac{1}{t-s} = \frac{1}{p-s} \left\{ \frac{1}{1-(t-p)/(s-p)} \right\} = \frac{1}{p-s} \sum_{k=0}^{\infty} \left(\frac{t-p}{s-p} \right)^k. \quad (32)$$

By combining (20) and (32) we obtain

$$\frac{1}{(1+t)^\beta(t-s)} = \sum_{k=0}^{\infty} B_k(s) \left(\frac{t-p}{s-p} \right)^k, \quad B_k(s) = \frac{1}{(1+p)^\beta(p-s)} \sum_{j=0}^k (-1)^j \frac{\Gamma(\beta+j)}{\Gamma(\beta)j!} \left(\frac{s-p}{1+p} \right)^j, \quad (33)$$

which reduces to (27) by multiplying with $(1-t)^{-\alpha}$ and by integrating, where

$$I_{i,k}(s) = \int_{t_i}^{t_{i+1}} (1-t)^{-\alpha} \left(\frac{t-p}{s-p} \right)^k dt. \quad (34)$$

The recursion in (28) can be easily deduced from the definition of $B_k(s)$ in (33). Moreover, since $\frac{|s-p|}{|1+p|} \leq 1$ for $p \in [0, 1]$ and $s \in (-1, 1)$, this recursion is numerically stable.

Equation (29) may be derived by dividing (17) by $(s-p)^k$. If $\frac{|1-p|}{|s-p|} < 1$ then (29) is numerically stable. If $\frac{|1-p|}{|s-p|} > 1$ then (29) is unstable, but its equivalent backward recursion is stable. To determine a starting point N for which the backward recursion computes $I_{i,k}(s)$, $k = (N-1)(1)0$ with an error less than ϵ , we use (34) to derive

$$|I_{i,N}(s)| \leq \frac{|t_{i+1}-t_i|^N}{|s-p|^N} |I_{i,0}| \leq \epsilon, \quad \text{implies } N \geq \frac{\ln(\epsilon/|I_{i,0}|)}{\ln(\lambda_2)}, \quad \lambda_2 = \frac{h_i}{|s-p|} \quad (35)$$

where $|I_{i,0}|$ is defined in (17) and $h_i = t_{i+1}-t_i$. By setting $I_{i,N}(s) = 0$, then the error ϵ will be multiplied by $\frac{|s-p|}{|1-p|} < 1$ in each recursion step and therefore all $I_{i,k}(s)$, $k = (N-1)(1)0$, will be computed with an error less than ϵ , provided that the maximum error in floating point arithmetic is less than ϵ [3, p. 17]. \square

We now consider the rate of convergence of the series in (27). We define

$$I_i^{(N)}(s) = \sum_{k=0}^N d_k(s), \quad d_k(s) = B_k(s) I_{i,k}(s), \quad R_N(s) = I_i(s) - I_i^{(N)}(s) = \sum_{k=N+1}^{\infty} d_k(s) \quad (36)$$

Theorem 5: The remainder $R_N(s)$ of the series in (27) satisfies:

$$|R_N(s)| \leq C(s) \lambda_2^{N+1} |I_{i,0}| \quad \text{for all } N \geq M = \left\lceil \frac{1-\beta}{2} \right\rceil, \quad (37)$$

where $\lambda_2 < 1$ is defined in (35) and $C(s)$ is independent of N .

Proof: We can easily see that the choice of M in (37) and $\beta < 1$, implies $|j-1+\beta| < j$ for all $j \geq M$. Thus, from the definition of $b_j(s)$ in (28) we derive

$$|b_j(s)| \leq \lambda_1^{j-L} |b_L(s)|, \quad j = L, L+1, \dots, \quad \text{for all } L \geq M \quad (38)$$

where $\lambda_1 = \frac{|s-p|}{1+p} < 1$ for $p \in [0, 1]$ and $s \in (-1, 1)$. The second equation in (33) implies

$$B_k(s) = B_\infty(s) - \sum_{j=k+1}^{\infty} b_j(s), \quad B_\infty(s) = \frac{(1+s)^{-\beta}}{p-s}, \quad (39)$$

and by using the sum of the geometric series and (38), we derive

$$|B_k(s)| \leq |B_\infty(s)| + \frac{|b_{k+1}(s)|}{1-\lambda_1}, \quad k \geq M. \quad (40)$$

Furthermore, (34) implies

$$|I_{i,k}(s)| \leq \lambda_2^{k-L} |I_{i,L}(s)|, \quad k = L, L+1, \dots, \quad \text{and } L \geq 0, \quad (41)$$

and since $|t_{i+1} - t_i| < |s-p|$ from the assumptions of Theorem 4, then $\lambda_2 < 1$. By substituting (41) and (40) into the last equation in (36) and by using (38) we derive

$$|R_N(s)| \leq \sum_{k=N+1}^{\infty} \left\{ |B_\infty(s)| \lambda_2^{k-N-1} + \lambda_1^{k-N} \frac{|b_{N+1}(s)|}{(1-\lambda_1)} \lambda_2^{k-N-1} \right\} |I_{i,N+1}| \leq \quad (42)$$

$$\left\{ \frac{|B_\infty(s)|}{(1-\lambda_2)} + \frac{|b_{N+1}(s)|}{(1-\lambda)\lambda_1(1-\lambda_1)} \right\} |I_{i,N+1}|,$$

where $\lambda = \lambda_1 \lambda_2 = \frac{h_i}{1+p}$ is also defined in Theorem 3. By using (41) the last inequality reduces to (37), where, because of (38), we may choose

$$C(s) = \left\{ \frac{|B_\infty(s)|}{(1-\lambda_2)} + \frac{|b_M(s)|}{(1-\lambda)(1-\lambda_1)} \right\}. \quad \square$$

We see that the speed of convergence of $R_N(s)$ depends on $\lambda_2 < 1$. If λ_2 is close to 1, then $R_N(s)$ will converge slowly. Since $s \in [t_j, t_{j+1}]$, slow convergence may be encountered in the intervals $[t_{j-1}, t_j]$ and $[t_{j+1}, t_{j+2}]$, particularly if s is near t_j or t_{j+1} . To avoid this, we select s to be a node point. Since $A_j(s)I_j(s) = 0$ and $A_{j-1}(s)I_{j-1}(s) = 0$ if $s = t_j$, then we do not have to evaluate $I_{j-1}(s)$ and $I_j(s)$ in these intervals. Furthermore, if we assume that the mesh is uniform, then we can easily see that $\lambda_2 = 1/2, 1/3, \dots$ in $[t_{j-2}, t_{j-1}]$, $[t_{j-3}, t_{j-2}]$, ..., respectively. Note that $(1/2)^{25} = 3 \times 10^{-8}$, $(1/3)^{16} = 7 \times 10^{-8}$, $(1/4)^{12} = 6 \times 10^{-8}$, which imply that only few subintervals near $[t_j, t_{j+1}]$ will need $N \geq 10$ for $R_N(s)$ to become less than $C(s) 10^{-8}$.

3. NUMERICAL APPLICATIONS

The algorithm described by Theorems 1, 2 and 4 is simple and computationally efficient. Note that the numerical computation of I_i and $I_i(s)$ requires the evaluation of a single summation as opposed to three or four nested summations needed by Miller and Keer's [13] algorithm. Moreover, for a uniform mesh, the total amount of computations needed for the evaluation of I_i and $I_i(s)$ may be cut in half by using the identity $\int_{t_i}^{t_{i+1}} = \int_{t_i}^{t_{i+2}} - \int_{t_{i+1}}^{t_{i+2}}$ where it is assumed that the I_i and $I_i(s)$ are saved (stored) when the mesh increases from $n+1$ to $2n+1$ points. We now present two examples which illustrate the advantages of the piecewise-polynomial quadratures versus the Gauss-Jacobi quadratures, particularly for "badly" behaved integrands.

Example 1: The piecewise-polynomial quadrature can be used to approximate weakly singular integrals of the form:

$$J = \int_{-1}^1 w(t) g(t) dt \approx J_n = \sum_{i=0}^n w_i g(t_i), \quad w(t) = (1-t)^{-\alpha}(1+t)^{-\beta} \quad (43)$$

$$w_0 = I_0 - \frac{I_0^{(1)}}{h_0}, \quad w_n = \frac{I_{n-1}^{(1)}}{h_{n-1}}, \quad w_i = I_i - \frac{I_i^{(1)}}{h_i} + \frac{I_{i-1}^{(1)}}{h_{i-1}}, \quad i = 1(1)(n-1), \quad (44)$$

where I_i is defined in Theorem 1 and $I_i^{(1)} = \int_{t_i}^{t_{i+1}} w(t) (t-t_i) dt$. The series expansion for $I_i^{(1)}$ may be obtained from (16) and (18) by replacing $I_{i,k}$ with $I_{i,k+1}$.

Extensions to higher order piecewise-polynomial approximations, i.e. quadratic, cubic, etc., is straightforward. The weights for piecewise-polynomial methods of order $m \geq 1$ may be obtained as linear combinations of $I_i^{(j)} = \int_{t_i}^{t_{i+1}} w(t) (t-t_i)^j dt$, $j = 0(1)m$. The integrals $I_i^{(j)}$ may be estimated by replacing $I_{i,k}$ with $I_{i,k+j}$ in (16) and (18). Since $I_{i,k+j}$ is used in the evaluation of all quadrature weights, very few additional computations are needed for the estimation of the weights of higher order methods.

In Table 3-1 we present the numerical results of the quadrature in (43) for $g(t) = \sqrt{|t|}$, $w(t) = (1-t^2)^{-1/2}$ and for two different choices of mesh points. The decay exponent and the order of convergence of the $Error = e_n = |J - J_n|$ are defined by $p = \ln(e_n/e_{2n})/\ln(2)$ and $O(n^{-p})$ respectively. The decay exponent for the uniform mesh $t_i = -1 + ih$, $i = 0(1)n$, $h = 2/n$ tends to $p = 1.50$, which is expected for the square root integrand (Atkinson [1, p. 255]). The global Gauss-Chebyshev quadrature, e.g. $J_g = \pi \sum_{j=0}^n \sqrt{|t_j|}/n$, $t_i = \cos((2i-1)\pi/(2n))$, $i = 1(1)n$, also converges with $p = 1.50$. The accuracy of the Gauss-Chebyshev quadrature is similar to the piecewise-linear quadrature

with uniform mesh, i.e. if $n = 80$, $J_g = 2.39723$. The value J_∞ is computed from the exact $J = 4\sqrt{\pi}\Gamma(3/4)/\Gamma(1/4)$.

It is well known that a better rate of convergence may be attained if the mesh concentrates at the points for which the input functions or their derivatives are singular. For the $\sqrt{|t|}$ whose derivative is singular at $t = 0$, the nonuniform mesh $t_i = -(1 - ih)^4$, $i = 0(1)(n/2)$, and $t_i = (-1 + ih)^4$, $i = (n/2)(1)n$, results in an $O(h^2)$ rate of convergence (column 7 of Table 3-1). The $O(h^2)$ rate is expected for this choice of nonuniform mesh (De Boor [4, p. 46]). However, a similar improvement on the rate of convergence of the Gauss-Chebyshev method is not possible since t_i are predetermined. For piecewise-polynomial quadratures, an additional extrapolation may be used to further improve the numerical results. This is demonstrated in the last column of Table 3-1, where $Extrap = (4J_{2n} - J_n)/3$. It can be shown that the order of convergence for the piecewise-linear quadrature in (43) is $O(h^2)$ for C^2 functions (Atkinson [1, p. 273]).

n	Uniform Mesh			Nonuniform Mesh			
	J_n	Error	p	J_n	Error	p	Extrap
10	2.36119	3.5×10^{-2}	---	2.36147	3.4×10^{-2}	---	----
20	2.38366	1.3×10^{-2}	1.47	2.38703	9.3×10^{-3}	1.91	2.39555
40	2.39176	4.5×10^{-3}	1.48	2.39388	2.4×10^{-3}	1.94	2.39616
80	2.39467	1.6×10^{-3}	1.49	2.39566	6.2×10^{-4}	1.96	2.39626
∞	2.39628		1.50	2.39628		2.00	2.39628

Table 3-1: The quadrature (43) for $g(t) = \sqrt{|t|}$ and $\alpha = \beta = 1/2$

Example 2: We will now apply the piecewise-polynomial quadrature to the solution of a CSIE with variable coefficients. Welstead [17, p. 103] has solved the CSIE

$$\sin\left(\frac{\pi s}{2}\right) w(s)y(s) + \frac{1}{\pi} \int_{-1}^1 \frac{\cos\left(\frac{\pi t}{2}\right) w(t) y(t)}{t-s} dt = f(s) \quad (45)$$

$$w(t) = (1-t)^{-1/2}(1+t)^{-1/2}\Omega(t), \quad \Omega(t) = e(1-t)^{t/2}(1+t)^{-t/2} \quad (46)$$

by using orthogonal polynomial approximations. Equation (45) is equivalent to

$$y(t) = \sin\left(\frac{\pi t}{2}\right) f(t) - \frac{1}{\pi} \int_{-1}^1 \frac{\cos\left(\frac{\pi s}{2}\right) f(s)}{w(t)(s-t)} ds + C, \quad (47)$$

where C is an arbitrary constant. By rewriting (47) as

$$y(t) = \sin\left(\frac{\pi t}{2}\right) f(t) - J(F; t) - F(t) \int_{-1}^1 \frac{\sqrt{1-t^2}}{s-t} ds + C, \quad (48)$$

$$F(t) = \frac{\cos\left(\frac{\pi t}{2}\right) f(t)}{e^{\pi \Omega(t)}}, \quad J(F; t) = \int_{-1}^1 \sqrt{1-s^2} \frac{F(s)-F(t)}{s-t} ds, \quad \int_{-1}^1 \frac{\sqrt{1-s^2}}{s-t} ds = -\pi t \quad (49)$$

then the piecewise-polynomial quadrature can be used to approximate $J(F; t)$.

In Table 3-2 we present the approximation $y_n(1)$ of $y(1)$ by the piecewise-linear quadrature (15) for $f(t) = \sqrt{1-t^2}$ and $C = 1/\pi$. We again select a uniform mesh $t_i = -1 + ih$, $i = 0(1)n$, $h = 2/n$ and a nonuniform mesh $t_i = -1 + (ih)^4$, $i = 0(1)(n/2)$, $t_i = -t_{n-i}$, $i = (n/2)(1)n$ which concentrates at ± 1 . Note that the decay exponent p tends to 2.00 for both choices of mesh points. This is expected, since

$$F(t) = (1-t)^{(1-t)/2} (1+t)^{(1+t)/2} \cos\left(\frac{\pi t}{2}\right) / (e^{\pi}) \quad (50)$$

is a $C^1[-1, 1]$ function and its second derivative is integrable. More specifically, we can show that $F(\pm 1) = 0$, $F'(\pm 1) = \mp 1/e$, $F''(t) = \Phi_1(t) \ln(1-t) + \Phi_2(t) \ln(1+t) + \Phi_3(t)$ where $\Phi_i(t)$, $i = 1(1)3$ are continuous functions. If $F_n(t)$ is the piecewise-polynomial approximation of $F(t)$ defined by (7), then, by using Peano's error formula, we easily see that

$$|J(F-F_n; 1)| \leq \frac{h^2}{8} \int_{-1}^1 (1+t)^{1/2} (1-t)^{-1/2} |F''(t)| dt \quad (51)$$

where $h = \max h_i$ for all i . Since $|\ln(1 \pm t)| < c(\delta)(1 \pm t)^{-\delta}$ for any $\delta > 0$, where $c(\delta)$ is a positive constant, the integral in the last equation is bounded and the $O(h^2)$ order of convergence is obvious.

We observe that the *Error* in the uniform mesh part of the Table is about half the *Error* of the nonuniform mesh, which is explained by (51) and the fact that $\max h_i$ is larger for the nonuniform mesh. However, the decay exponent p for the nonuniform mesh tends faster to 2.00 and consequently, the error for the nonuniform extrapolation is smaller than that of the uniform extrapolation for the same n . The $y_\infty(1) = 0.518592$ is computed by choosing $n = 2560$.

The numerical solution derived by Welstead [17], with the use of orthogonal polynomials with respect to $w(t)$ in (45), is $y_n(1) = 0.518583$ for $n = 40$. The generation of such polynomials, mesh points and quadrature weights for nonclassical weight functions require considerable computational effort since a Stieltjes procedure to determine the recurrence coefficients, Fejer's quadrature to evaluate the integrals involved and the solution of an eigenvalue problem via Lanczos' algorithm would have to be implemented (e.g. Welstead [17, p. 115]). Therefore, it is obvious that the computational effort will increase with the dimension n of the approximation. It is not clear whether this approach is computationally feasible for large n 's. However, the computational effort for the piecewise-polynomial method increases slowly as n increases, since the series for all additional integrals converge fast (see Theorems 3 and 5). Using the piecewise-linear method, we have routinely solved problems with $n = 5000$ without encountering any numerical instability.

n	Uniform Mesh			Nonuniform Mesh			
	$y_n(1)$	Error	p	$y_n(1)$	Error	p	Extrap
10	0.514213	4.3×10^{-3}	---	0.508712	9.9×10^{-3}	---	---
20	0.517349	1.2×10^{-3}	1.82	0.515989	2.6×10^{-3}	1.92	0.518414
40	0.518251	3.4×10^{-4}	1.87	0.517931	6.6×10^{-4}	1.98	0.518579
80	0.518501	9.0×10^{-5}	1.91	0.518426	1.7×10^{-4}	2.00	0.518592
∞	0.518592		2.00	0.518592		2.00	0.518591

Table 3-2: The quadrature (15) with $\alpha = \beta = -1/2$ and $g(t)$ defined in (50).

Note : All computations were performed on a DECSYSTEM/2060T using FORTRAN 77 with double precision floating point arithmetic (with a mantissa of 16 to 18 decimals and with an exponent in the range 0.14×10^{-38} to 3.4×10^{38}).

4. REFERENCES

1. K. E. Atkinson, *An introduction to numerical analysis* John Wiley & Sons, Inc, New York, 1978.
2. K. E. Atkinson, *The numerical evaluation of singular integrals of Cauchy type* SIAM J. Numer. Anal., 9 (1972), pp. 284-299.
3. G. Dahlquist & A. Bjorck, *Numerical Methods*, Prentice-Hall, New Jersey, 1974.
4. C. De Boor, *A practical guide to splines*, Applied Mathematical Sciences, 27, Springer-Verlag, New York, 1978
5. D. Elliott, *Orthogonal polynomials associated with singular integral equations having a Cauchy kernel*, SIAM J. Math. Anal., 13 (1982), pp. 1041-1052
6. F. Erdogan, G.D. Gupta and T.S. Cook, *Numerical solution of singular integral equations*, Mechanics of Fracture, Vol. I, (1973), pp. 368-425.
7. A. Gerasoulis, *Piecewise polynomial approximations in the solution of singular integral equations*, Advances in Computer Methods for Partial Differential Equations, IMACS, Rutgers U., 4 (1981), pp. 386-390.
8. A. Gerasoulis, *The use of piecewise quadratic polynomials for the solution of singular integral equations of Cauchy type*, Comp. and Maths. with Appls an Int. J., 8 (1982), pp. 15-22.
9. A. Gerasoulis and R. Srivastav, *A method for the numerical solution of singular integral equations with a principal value integral*, Int. J. of Engng. Sci., 19 (1981), pp. 1293-1298.
10. V. V. Ivanov, *The theory of approximate methods and their applications to the numerical solution of singular integral equations*, Noordhoff, Leyden, 1976.
11. E. Jen, and R.P. Srivastav, *Cubic splines and approximate solution of singular integral equations*, Mathematics of Computation, (1981), p. 37.
12. Flugge-Lotz I., *Mathematical improvement of method for computing Poisson integrals involved in determination of velocity distribution on airfoils*, NACA, Report 2451, 1951.
13. G. R. Miller and L. M. Keer, *A numerical technique for the solution of singular integral equations of the second kind*, Quarterly of Applied Mathematics, (1985), pp. 455-465.
14. D. F. Paget & D. Elliott, *An Algorithm for the numerical evaluation of certain Cauchy principal value integrals*, Mathematics of Computation, 19 (1972), pp. 373-385
15. C. Stewart, *On the numerical evaluation of singular integrals of cauchy type* J. Soc. Indust. Appl. Math. (SIAM), 8 (1960), pp. 342-353.
16. F. G. Tricomi, *On the finite Hilbert transform*, Quart. J. of Math., 2 (1951), pp. 199-211.
17. S. Welstead, *Orthogonal polynomials applied to the solution of singular integral equations*, PH.D. Thesis, Purdue University, 1982

NUMERICAL SOLUTION OF RANDOM LINEAR
VOLTERRA INTEGRAL EQUATION*

M. Sambandham
Department of Mathematics
Atlanta University/Morehouse College
Atlanta, GA 30314

ABSTRACT. Numerical solutions are obtained to Volterra integral equations with random nonhomogeneous terms. The method we use is Lobatto quadrature formula. Based on the simulation of random forcing term the numerical solutions are used (i) to compare the convergence of the average of the random solutions to the solution of the average equation, (ii) to evaluate the confinement probability and risk functionals, and (iii) to discuss the convergence of the solution processes of the random equations to the solution of the deterministic equation through the sample path graphs, as the variance of the random forcing term tends to zero.

I. INTRODUCTION. An integral equation of the form

$$y(x) = g(x) + \int_{x_0}^x K(x,t) y(t) dt \quad (1)$$

is said to be a Volterra integral equation of the second kind in which $K(x,y)$ is the kernel and $g(x)$ is the nonhomogeneous term (or forcing term). Probabilistic analog of (1), namely, a random Volterra integral equation is defined as follows:

$$y(x,\omega) = g(x,\omega) + \int_{x_0}^x K(x,s,\omega) y(s,\omega) ds, \quad (2)$$

where $K(x,y,\omega)$ is the random kernel and $g(x,\omega)$ is the random forcing term. The parameter ω is an element of a given probability measure space $(\Omega, \mathcal{A}, \mu)$.

In this article we obtain the numerical solution of (2). We take the available deterministic methods of solving a Volterra integral equation and suitably adopt these methods to generate the sample solutions of the random Volterra integral equation (2).

*Research supported by ARO Grant No. DAAG29-85-G0109.

For related work on random Fredholm equation of second kind see [3,9]. In [3] random Fredholm equation is considered with a random forcing term or a random kernel. In [9] a singular integral equation with random forcing term is considered. Lax [8] has used the method of moments to obtain the mean and the autocorrelation of random Volterra integral equation. Tsokos and Padgett [10] have employed the method of successive approximation to solve random nonlinear Volterra integral equations. For the several discussion and methods of numerical solution of Volterra integral equations refer to Baker [1], Golberg [4]. Other important numerical method is Lobatto Method (Jain and Sharma [7]).

We organize our article as follows. In Section 2 we present a short review of the Lobatto method [7]. In Section 3 we present the numerical technique we employed to solve the random Volterra integration equation. The concept of confinement probability and risk functionals are discussed in Section 4. Some numerical results and figures are presented in Section 5 and a short discussion and conclusion are in Section 6.

II. LOBATTO QUADRATURE FORMULA. Consider the Volterra linear integral equation of second kind:

$$y(x) = g(x) + \int_{x_0}^x K(x,t)y(t)dt. \quad (3)$$

In (3) $y(x)$ is numerically to be determined for a given continuous forcing function $g(x)$ and jointly continuous kernel $K(x,t)$. Equation (3) gives the values of $y(x)$ at $x_n = x_0 + nh$ as

$$\begin{aligned} y(x_n) &= g(x_n) + \int_{x_0}^{x_n} K(x_n,t)y(t)dt \\ &= g(x_n) + \sum_{p=0}^{n-1} \int_{x_p}^{x_{p+1}} K(x_n,t)y(t)dt. \end{aligned} \quad (4)$$

Now approximating the integral in (4) by the following quadrature formula

$$\int_{x_0}^{x_0+h} g(x)dx = \frac{h}{2} \sum_{p=1}^4 W_p g(\tau_p), \quad (5)$$

where

$$\begin{aligned}
W_1 &= W_4 = 1/6, & W_2 &= W_3 = 5/6 \\
\tau_1 &= x_0, \\
\tau_2 &= x_0 + rh, & r &= (5 - \sqrt{5})/10 \\
\tau_3 &= x_0 + sh, & s &= (5 + \sqrt{5})/10 \\
\tau_4 &= x_0 + h,
\end{aligned}$$

$$R = \frac{-4h^7 g^{(6)}(t)}{3.2^7 \cdot 15750}, \quad x_0 < t < x_0 + h,$$

the approximate value of $y(x_n)$ is given by

$$\begin{aligned}
y(x_n) &= [12g(x_n) + h\{K(x_n, x_{n-1})y(x_{n-1}) \\
&\quad + 5K(x_n, x_{n+r-1})y(x_{n+r-1}) \\
&\quad + 5K(x_n, x_{n+s-1})y(x_{n+s-1})\} \\
&\quad + h \sum_{p=0}^{n-2} \{K(x_n, x_p)y(x_p) \\
&\quad + 5K(x_n, x_{p+r})y(x_{p+r}) \\
&\quad + 5K(x_n, x_{p+s})y(x_{p+s}) \\
&\quad + K(x_n, x_{p+1})y(x_{p+1})\}] / (12 - hK(x_n, x_n)),
\end{aligned} \tag{6}$$

where

$$\begin{aligned}
y(x_{p+r}) &= \frac{\gamma h}{\Delta \sqrt{5}} y(x_p) (h\beta_{1p} - \beta_{2p}) - \frac{\gamma^2}{\Delta} y(x_p) \\
&\quad + [(\alpha_{2p}\beta_{0p} - \alpha_{0p}\beta_{2p})y(x_p) + \alpha_{2p}g(x_{p+s}) \\
&\quad - \beta_{2p}g(x_{p+r})] [\frac{h^3}{5\sqrt{5}} \beta_{1p} - \gamma_{sh}] / \Delta \\
&\quad - [(\alpha_{1p}\beta_{0p} - \alpha_{0p}\beta_{1p})y(x_p) + \alpha_{1p}g(x_{p+s}) \\
&\quad - \beta_{1p}g(x_{p+r})] [\frac{h^3}{5\sqrt{5}} \beta_{2p} - \gamma s^2 h^2] / \Delta,
\end{aligned}$$

$$\begin{aligned}
y(x_{p+s}) = & \frac{\gamma h}{\Delta \sqrt{5}} y(x_p) (h\alpha_{1p} - \alpha_{2p}) - \frac{\gamma}{\Delta} y(x_p) \\
& + [(\alpha_{2p}\beta_{0p} - \alpha_{0p}\beta_{2p})y(x_p) + \alpha_{2p}g(x_{p+s}) \\
& - \beta_{2p}g(x_{p+r})] [\frac{h^3}{5\sqrt{5}} \alpha_{1p} - \gamma rh] / \Delta \\
& - [(\alpha_{1p}\beta_{0p} - \alpha_{0p}\beta_{1p})y(x_p) + \alpha_{1p}g(x_{p+s}) \\
& - \beta_{1p}g(x_{p+r})] [\frac{h^3}{5\sqrt{5}} \alpha_{2p} - \gamma r^2 h^2] / \Delta,
\end{aligned}$$

$$\alpha_{np} = \int_{x_0}^{x_{p+r}} K(x_{p+r}, x) (x - x_p)^n dx, \quad n = 0, 1, 2,$$

$$\beta_{np} = \int_{x_0}^{x_{p+s}} K(x_{p+s}, x) (x - x_p)^n dx, \quad n = 0, 1, 2,$$

$$\gamma = \alpha_{1p}\beta_{2p} - \alpha_{2p}\beta_{1p},$$

$$\Delta = \gamma [h(\alpha_{1p}s^2h - \beta_{1p}r^2h + r\beta_{2p} - s\alpha_{2p}) - \frac{h^3}{5\sqrt{5}} - \gamma].$$

This equation (6) describes a single-step method for the solution of (3).

III. RANDOM VOLTERRA INTEGRAL EQUATION AND THE NUMERICAL METHOD. The equation that we are interested in is the following equation:

$$y(x, \omega) = g(x, \omega) + \int_{x_0}^x K(x, s, \omega) y(s, \omega) ds \quad (7)$$

where $g(x, \omega)$ is the random forcing function to be simulated. For each sample function $g(x, \omega)$ using Lobatto method equation (7) is solved. For our numerical work we consider the following two examples:

Example 1: $g_1(x, \omega) \in N(x, \sigma^2 x)$,
 $K_1(x, s, \omega) = s - x$,
 $\sigma^2 = \text{constant}$,
 $x_0 = 0$.

Example 2: $g_2(x, \omega) \in N(x+1, \sigma^2 x)$
 $K_2(x, s, \omega) = x - s,$
 $\sigma^2 = \text{constant}, x_0 = 0.$

We notice that in Examples 1 and 2 $Eg_1(x, \omega) = x$ and $E(g_2(x, \omega)) = x+1$. These factors show that the solutions of the average of equation (7) for Examples 1 and 2 are respectively $\sin x$ and e^x . Further we note that if $\sigma = 0$ then the random equation (7) becomes deterministic. For our numerical calculations we take σ values tend to zero so that we get the deterministic solution at $\sigma = 0$. This factor is an additional advantage that we get the deterministic solution as a particular case of our numerical example. The graphs and numerical data for different σ are presented in Section 5.

IV. CONFINEMENT PROBABILITY AND RISK FUNCTIONAL. The concept of confinement probability and risk functional are useful techniques in the probabilistic numerical analysis. To be more specific these are useful in the analysis of stability of continuous random systems with nonstationary responses. The confinement probability can be considered to be a real valued functional of random function. For any positive constant α , the confinement probability function F_α for any random solution $y(x)$ is defined by

$$F_\alpha(y) = P(|y(x)| \leq \alpha, \text{ for all } x \in [0, t]) \quad (8)$$

Those systems for which the confinement probability decays to zero slowly are considered to be more stable than those for which the confinement probability approaches to zero more rapidly as the system evolves in time.

The concept of risk functional is useful to check whether the system exceeds certain bounds. Suppose there is certain risk if the absolute value of y exceeds certain constant γ . Then the risk functional G_γ is defined by the relation

$$G_\gamma(y) = P\{[\min_{0 \leq \xi \leq x} y(\xi) \leq -\gamma] \cup [\max_{0 \leq \xi \leq x} y(\xi) \geq \gamma]\}. \quad (9)$$

Confinement probability and risk functional are useful techniques one can implement and analyze during the numerical procedure of probabilistic numerical analysis. These concepts can be used to check the stability and error analysis during the numerical experiments. In the next section, we discuss

these concepts for Examples 1 and 2. For more details on confinement probability and risk functionals see [5,6].

V. NUMERICAL RESULTS. From IMSL subroutine we generate standard normal random variables. Let it be $R(\omega)$. Then for $\sigma \geq 0$, let

- (i) $g_1(x, \omega) = x + \sigma\sqrt{x} R(\omega)$,
- (ii) $g_2(x, \omega) = x + 1 + \sigma\sqrt{x} R(\omega)$.

We remark that $E(g_1(x, \omega)) = x$ and $E(g_2(x, \omega)) = x+1$ and $\text{Var}(g_i(x, \omega)) = \sigma^2 x$, $i = 1, 2$. We notice that if $\sigma = 0$, $g_1(x, \omega)$ and $g_2(x, \omega)$ are deterministic functions. We simulate sample functions for $g_1(x, \omega)$ and $g_2(x, \omega)$ and use Lobatto method to solve Examples 1 and 2. Our numerical results are based on 40 samples. For Example 1 we analyze when $x \in [0, \pi]$ and for Example 2 we study for $x \in [0, 3]$. For the numerical studies we took $\sigma = 1.0, 0.5, 0.1, 0.05, 0.01, 0$. For these six values of σ , we present the sample path graphs for Examples 1 and 2 in the Figures 1-6 and 7-12 respectively. These graphs illustrate the convergence of the sample paths to the exact values ($\sigma = 0$). Table I and II represent the confinement probability and risk functional values for $\sigma = 1.0, \sigma = 0.1$ respectively for the Examples 1 and 2.

Table I.

$y(x) = \sin x, 0 \leq \sin x \leq 1$				
	Confinement Probability		Risk Functional	
x	$\sigma = 1.0$	$\sigma = 0.1$	$\sigma = 1.0$	$\sigma = 0.1$
(0, .2)	0.710	1.000	0.290	0.000
(0, .4)	0.675	1.000	0.325	0.000
(0, .6)	0.640	1.000	0.360	0.000
(0, .8)	0.610	1.000	0.389	0.000
(0, 1.0)	0.585	1.000	0.415	0.000
(0, 1.2)	0.567	0.984	0.433	0.016
(0, 1.4)	0.551	0.944	0.449	0.056
(0, π)	0.528	0.880	0.472	0.120

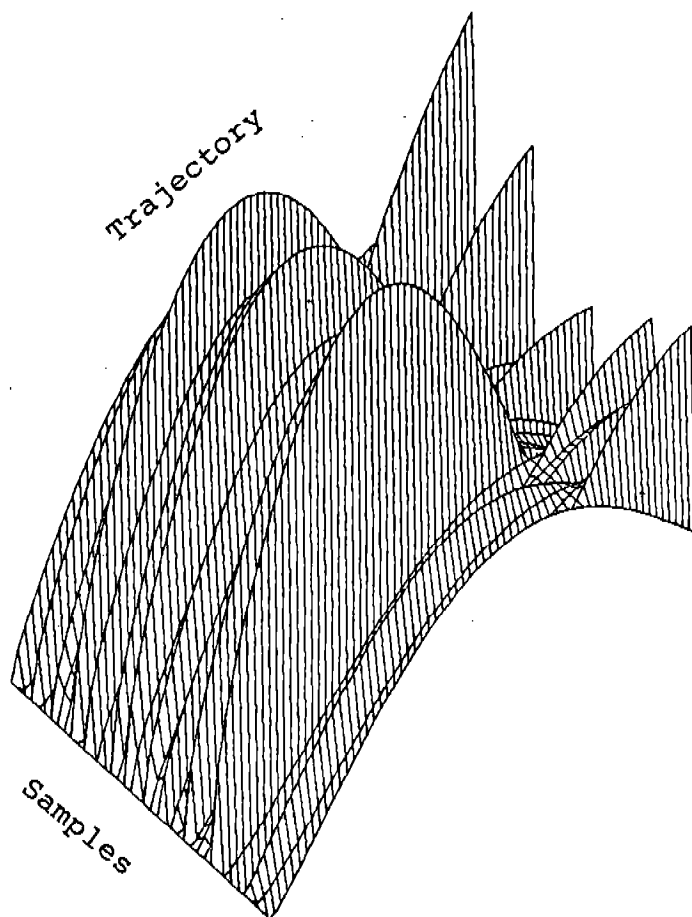


Fig 1 : $\sigma = 1.0$ (Example 1)

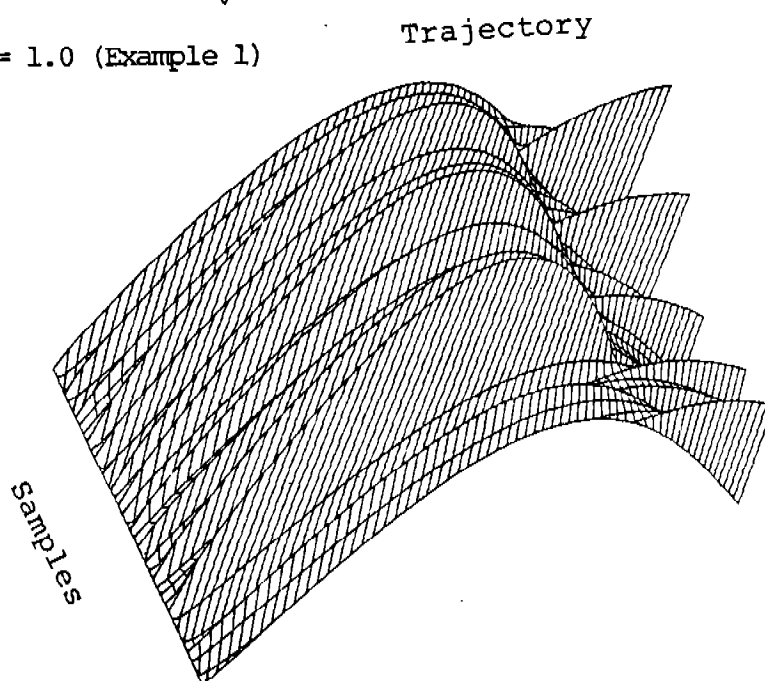


Fig 2 : $\sigma = 0.5$ (Example 1)

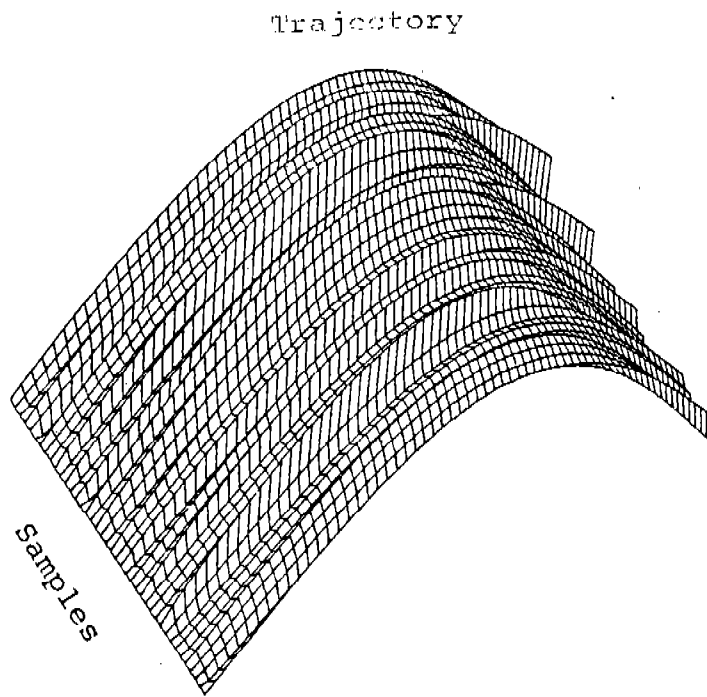


Fig 3 : $\sigma = 0.1$ (Example 1)

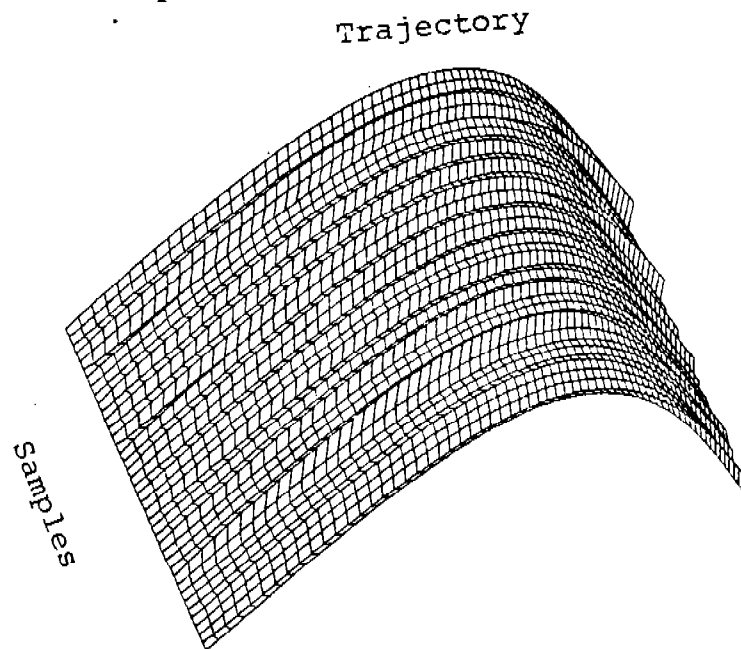


Fig 4 : $\sigma = 0.05$ (Example 1)

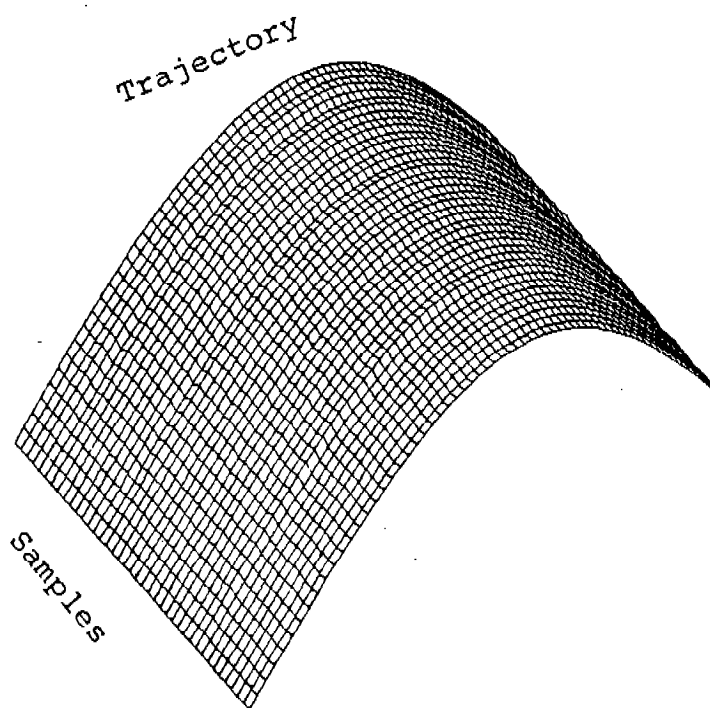


Fig 5: $\sigma = 0.01$ (Example 1)

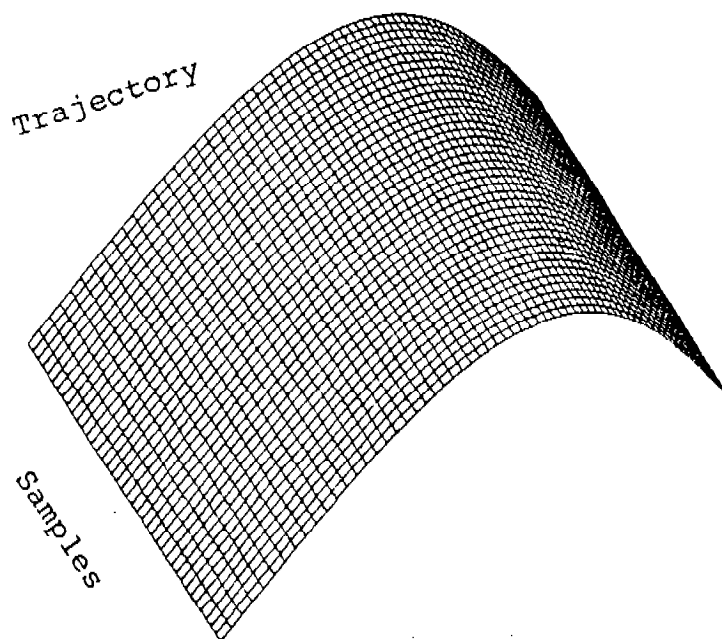


Fig 6: $\sigma = 0.0$ (Example 1)

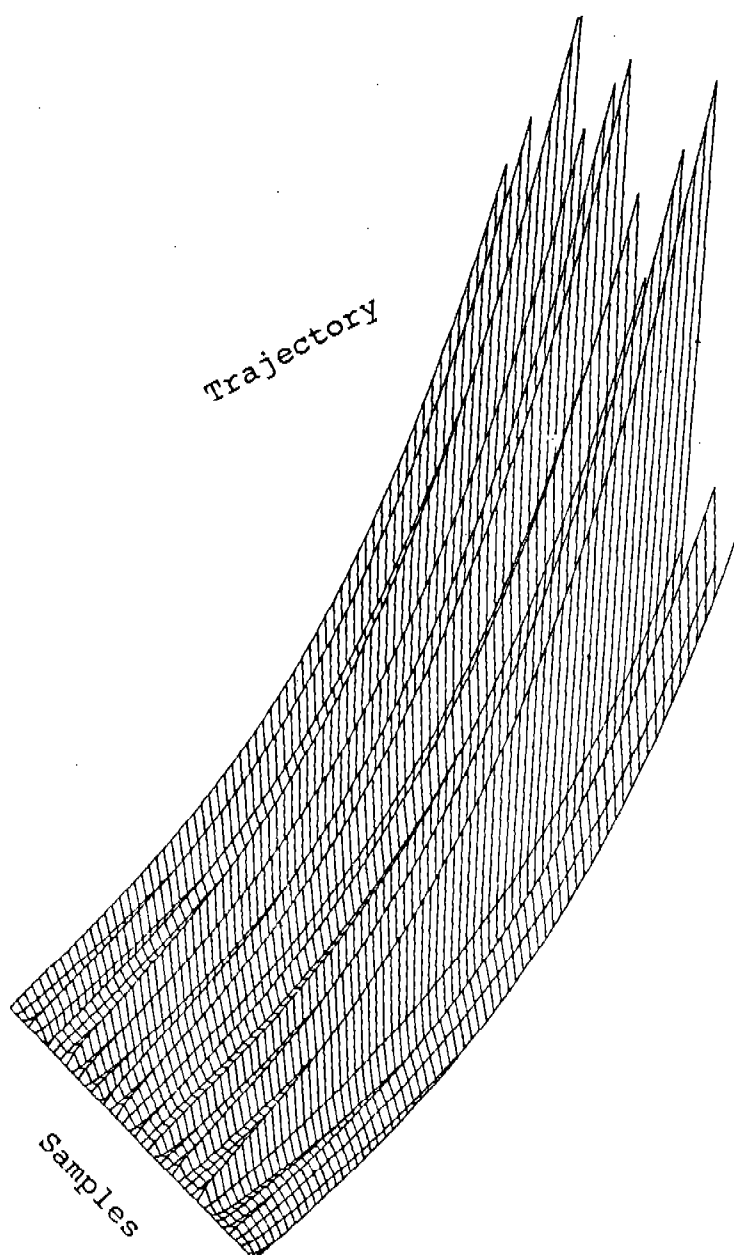


Fig 7: $\sigma = 1.0$ (Example 2)

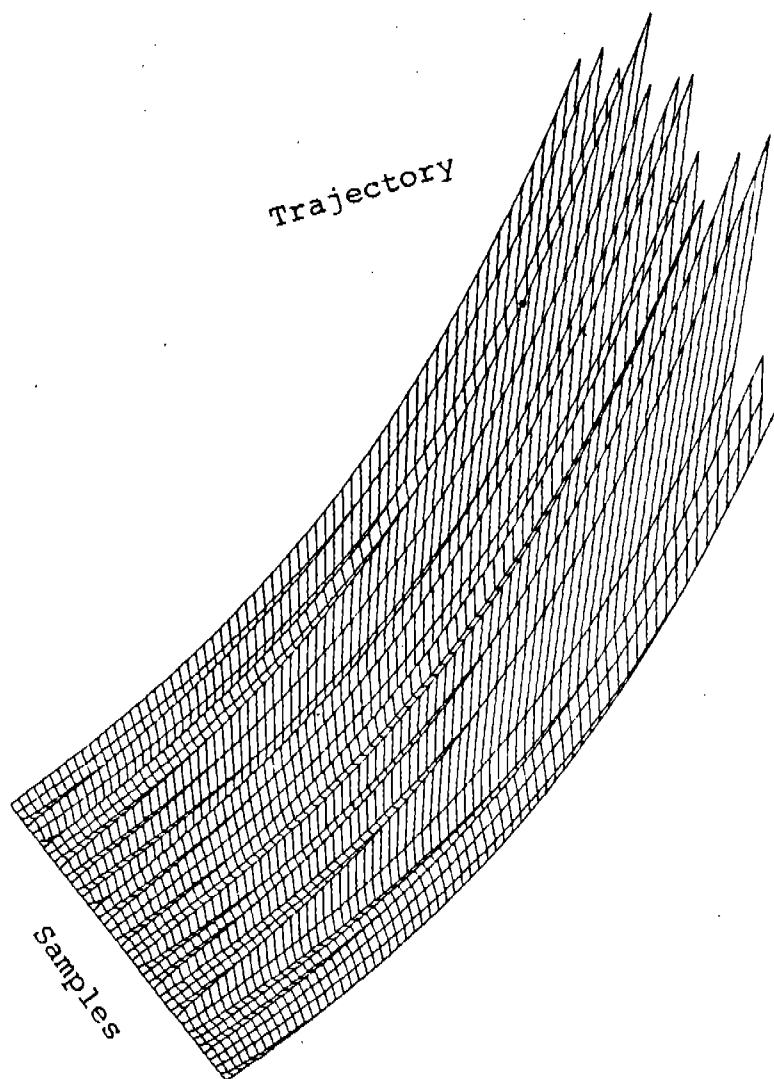


Fig 8: $\sigma = 0.5$ (Example 2)

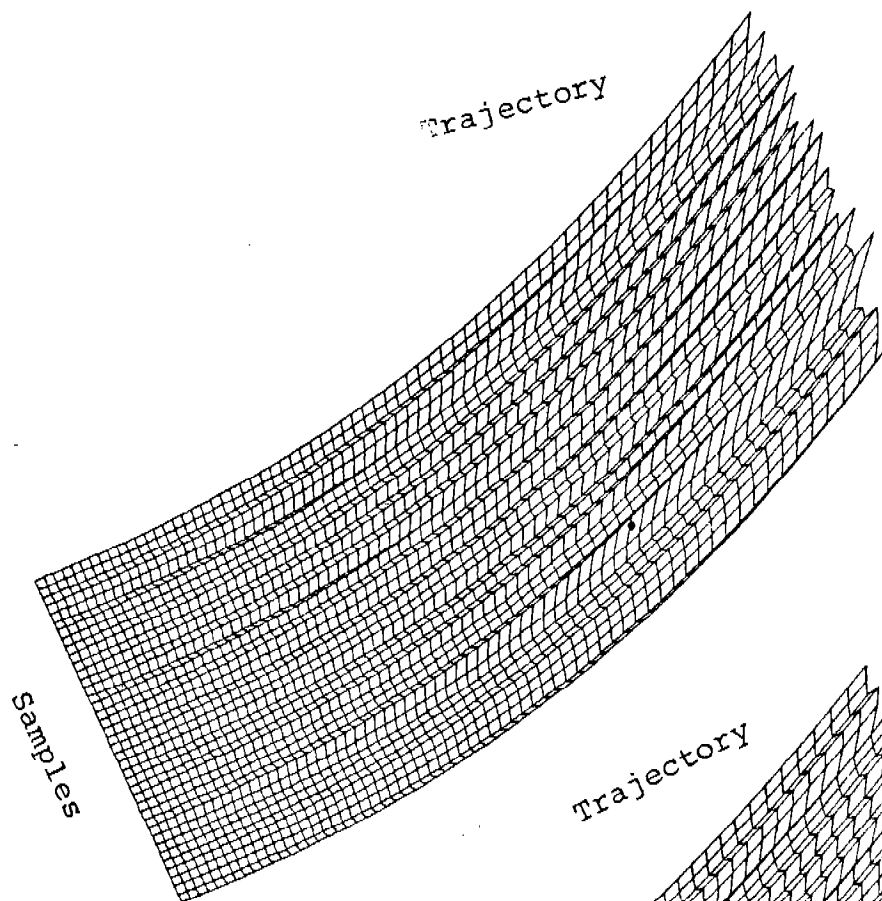


Fig 9: $\sigma = 0.1$ (Example 2)

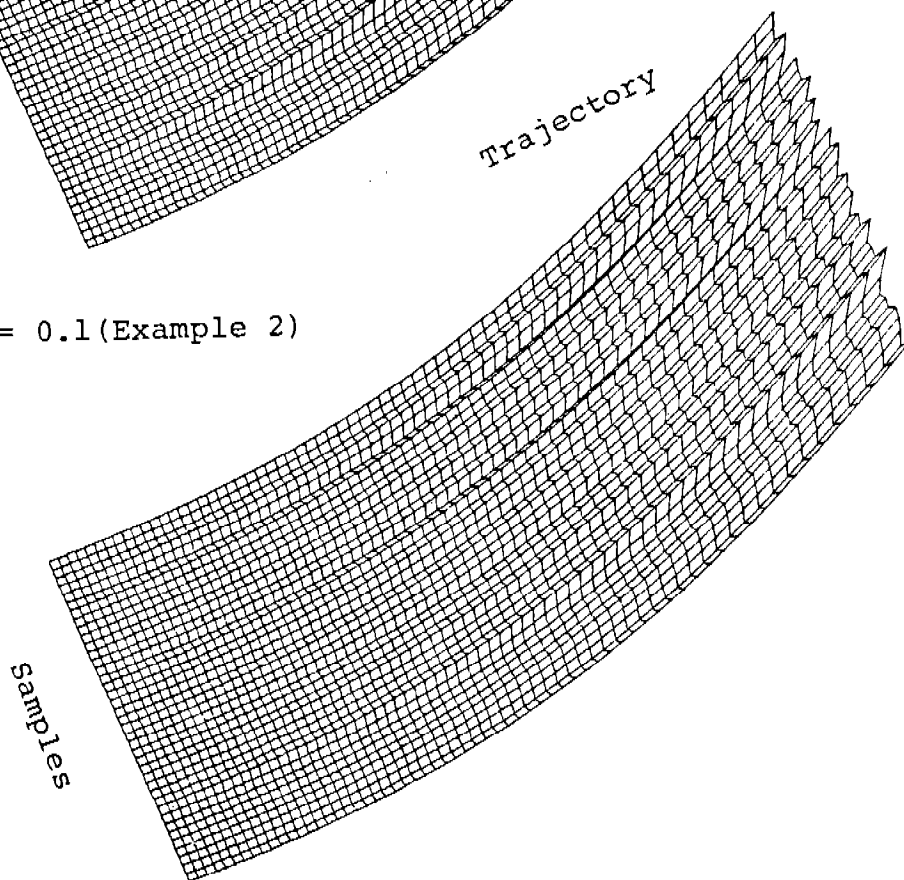


Fig 10: $\sigma = 0.05$ (Example 2)

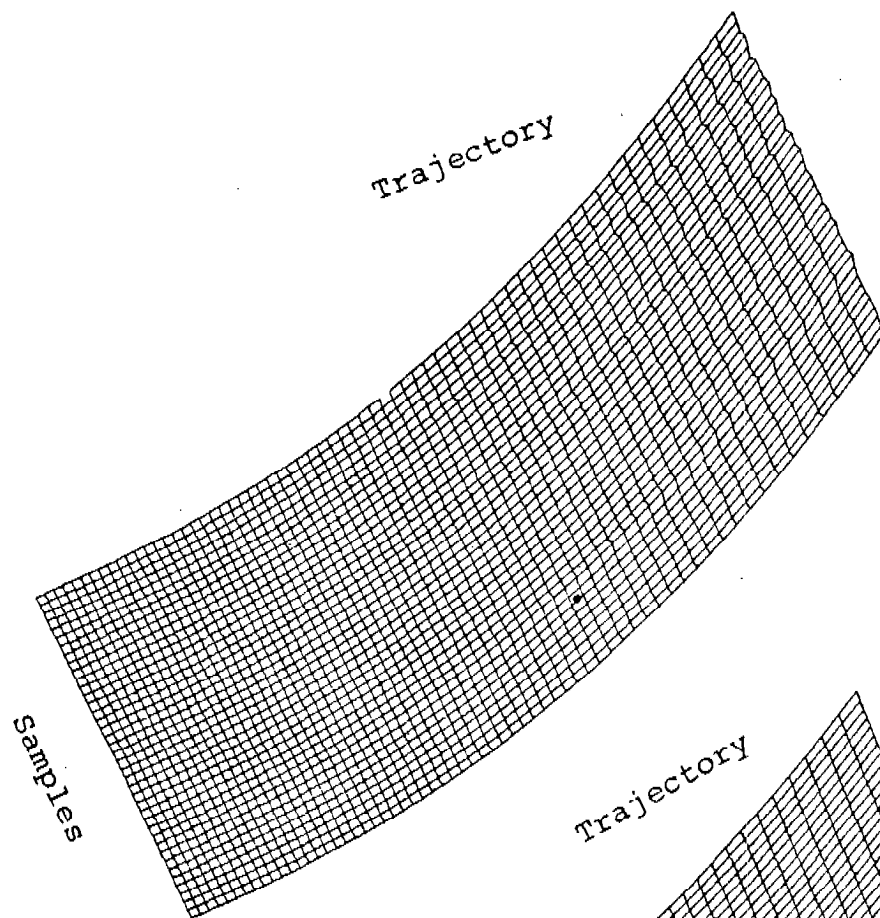


Fig 11: $\sigma = 0.01$ (Example 2)

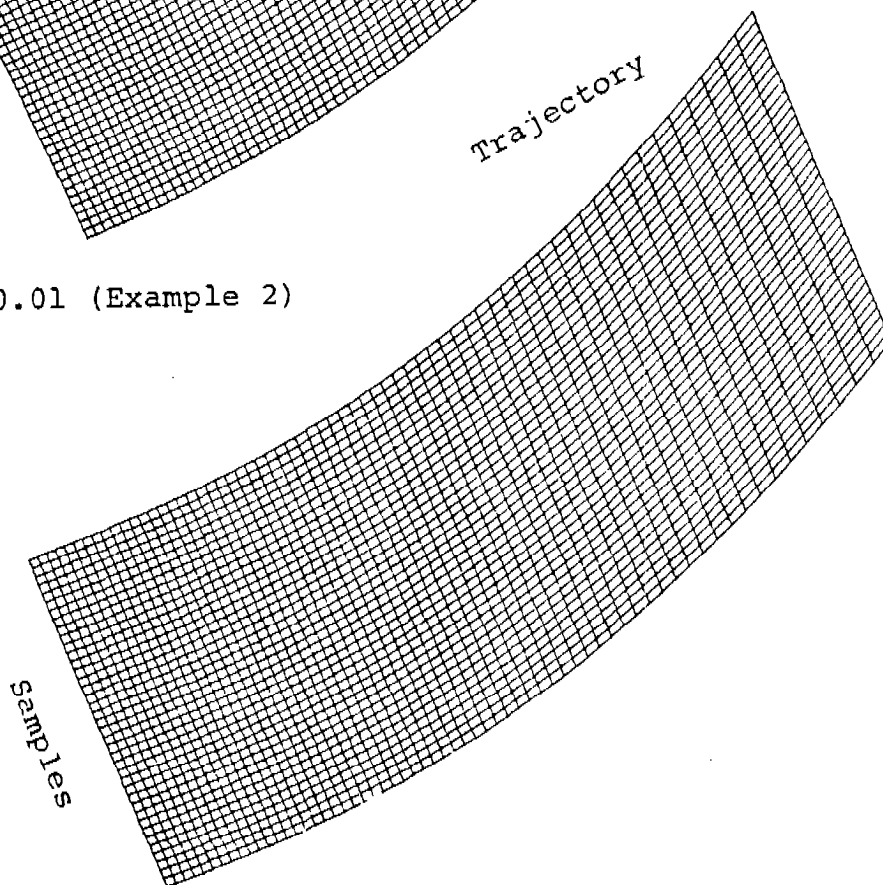


Fig 12 : $\sigma = 0.0$ (Example 2)

Table II.

$y(x) = e^x, \quad 1 < e^x < e^3$				
	Confinement Probability		Risk Functionals	
x	$\sigma = 1.0$	$\sigma = 0.1$	$\sigma = 1.0$	$\sigma = 0.1$
(0, .5)	0.738	1.000	0.262	0.000
(0, 1.0)	0.800	1.000	0.200	0.000
(0, 1.5)	0.837	1.000	0.163	0.000
(0, 2.0)	0.865	1.000	0.135	0.000
(0, 2.5)	0.887	1.000	0.113	0.000
(0, 3.0)	0.894	0.9901	0.106	0.099

VI. DISCUSSION. Figures 1-6 and 7-12 illustrate the convergence of the exact value. These figures clearly illustrate the random disturbances at each stage. Further the importance of the probabilistic models and deterministic models are clearly visualized from the graphs.

The discussion in Section 4 illustrate that confinement probability and risk functionals are useful techniques one can use in the probabilistic numerical analysis. The data in Tables I and II confirm that the random solution of the random equation converges faster to the solution $y(x) = \sin x$ of the corresponding deterministic equation in the case of $\sigma = 0.1$ compared to the case of $\sigma = 1.0$. For the same parameter values $\sigma = 0.1$ path converges in the second example

($y(x) = e^x$) even faster than the case of Example 1 ($y = \sin x$). Further the risk functionals illustrate the probability that the solution exceeded certain limits. Though these factors are utilized here to discuss the solution processes, one can suitably use these factors for analysis and other topics of consideration in the probabilistic numerical analysis.

REFERENCES

1. C. T. H. Baker. The Numerical Treatment of Integral Equations, Clarendon Press, Oxford, 1977.
2. A. T. Bharucha-Reid and M. J. Christensen. Approximate solution of random integral equations. Math. Comp. Simulation. 24 (1984), 321-328.
3. M. J. Christensen and A. T. Bharucha-Reid. Numerical solution of random integral equation I, II. J. Integral Equation. 3 (1981), 217-229, 333-344.
4. M. A. Golberg (editor). Solution Methods for Integral Equations. Theory and Applications. Plenum Press, New York and London 1979.
5. K. Gopalsamy and A. T. Bharucha-Reid. On a class of parabolic partial differential equations driven by stochastic point processes. J. Appl. Probl. 12 (1975), 98-106.
6. K. Gopalsamy. On the conditional density of noise driven evolution. ZAMM 56 (1976), 239-242.
7. M. Jain and K. Sharma. Numerical solution of linear differential equation and linear Volterra's integral equations using Lobatto method. The Computer Journal 10 (1967), 101-107.
8. M. Lax. Solving random linear Volterra integral equations using the methods of moments. J. Integral Equation. 3 (1981), 357-363.
9. M. Sambandham, M. J. Christensen and A. T. Bharucha-Reid. Numerical solution of random integral equation III. Random Chebyshev polynomials and Fredholm equations of the second kind, to be published.
10. C. P. Tsokes and W. J. Padgett. Random integral equations with applications to life sciences and engineering. Academic Press, New York, 1974.

COMMENTS ON FINITE ELEMENT METHOD AND BAND-WIDTH REDUCTION

WITH REFERENCE TO TRANSIENT HEAT CONDUCTION

R. Yalamanchili

Technology Branch, Armament Division
Fire Control and Small Caliber Weapon Systems Laboratory
US Army ARDC, AMCCOM, Dover, NJ 07801-5001

ABSTRACT. Transient two-dimensional heat conduction was analyzed by finite-element techniques (FEM). The most common assumptions such as linear temperature distribution within each element and within each time increment were utilized. The resulting difference equations were studied against the method of weighted residuals (MWR) and various finite-difference techniques (FDM) with emphasis on stability, accuracy and nonoscillation characteristics. These codes possess built-in algorithms, such as, the Cuthill-McKee algorithm for optimization of the bandwidth of the matrix. A ladle of molten metal was simulated by this algorithm. Even though it took 10 iterations, a minimum bandwidth of 26 was obtained. The manual numbering of nodes yielded the bandwidth as 17. There is a wide scope for further research and development of new and improved algorithms.

I. INTRODUCTION. The finite-element method of describing continuous systems was first introduced in the mid-1950's and has since become an extremely useful engineering technique. The FEM has been applied in a variety of fields including stress analysis, fluid dynamics, and field theory. Norrie and de-Vries [1] list over 7,000 references in a bibliography published in 1976. Numerous references exist in the case of the finite-difference method and also the method of weighted residuals. It is not the objective to discuss such a vast literature. Rather, it is the intent to summarize conclusions based on the author's research even though it may be considered controversial. Hopefully, such a statement may stimulate further research and bear fruitful findings later.

The transient two-dimensional heat conduction was analyzed by FEM with popular assumptions such as linear temperature distribution within each element and within each time-increment. The resulting equations were compared against the result of the MWR and FDM with emphasis on stability, accuracy, and nonoscillation characteristics. The effect of relaxation of those assumptions upon the conclusions is questioned.

Various FEM codes [2, 3, 4] are available for analysis of engineering problems with any geometry, nonlinear material properties, and nonlinear conditions. These codes possess algorithms, such as the Cuthill-McKee [5] algorithm, for optimization (minimization) of the bandwidth of the matrix (resulting system of equations). It is important to realize the effect of bandwidth on core storage and computational times, especially for multi-dimensional problems. It is demonstrated that these algorithms do not necessarily yield the minimum bandwidth and therefore, manual numbering of nodes may be better if practicable.

2. FINITE ELEMENT METHODOLOGY. Gurtin [6] introduced the variational principle for linear initial value problems in 1964 and confirmed that the function $T(x, y, t)$ which leads to an extremum of the functional

$$\Omega_1(T) = \frac{1}{2} \int_V [\rho C_p T^* T' + \nabla T^* K \nabla T' - 2\rho C_p T_0^* T'] dV - \int_S \hat{Q}_1 \eta_1^* T' dS \quad (1)$$

is the solution of the transient heat conduction equation

$$(K^* T, t)_t - \rho C_p^* \frac{\partial T'}{\partial t} = 0 \quad (2)$$

With the boundary condition

$$K^* T, t - \hat{Q}_1 = 0 \quad (3)$$

Where $T(x, y, t)$ is the temperature at the spatial point (x, y) and time t , T_0 is the initial temperature, ∇T is the gradient of T , K is the thermal conductivity, ρ is the material density, C_p is the heat capacity of the material per unit mass, $\hat{Q}_1(x, y, t) = \int Q_1(x, y, t) dt$. V is the volume and $*$ is the convolution symbol defined as

$$T^* T = \int_0^t T(x, y, t - \tau) T(x, y, \tau) d\tau$$

$$\nabla T^* \nabla T = \frac{\partial T^*}{\partial x} \frac{\partial T}{\partial x} + \frac{\partial T^*}{\partial y} \frac{\partial T}{\partial y} \quad (4)$$

The two-dimensional body was divided into square elements of length ΔL and linear temperature distribution was assumed within each element. The integration in the functional equation was accomplished and applied, the first variation with respect to the nodal temperature, $T_{i,j,k+1}$. An additional assumption regarding variation of nodal temperature with respect to time within any time increment is necessary in order to evaluate the mid-term of the functional. If a linear variation (with respect to time within each time step) is assumed, the finite-element difference equation (FEDE) can be written as:

$$\begin{aligned} & AT_{i-1,j-1,k+1} + BT_{i-1,j,k+1} + AT_{i-1,j+1,k+1} + BT_{i,j-1,k+1} \\ & + CT_{i,j,k+1} + BT_{i,j+1,k+1} + AT_{i+1,j-1,k+1} + BT_{i+1,j,k+1} \\ & + AT_{i+1,j+1,k+1} = DT_{i-1,j-1,k} + ET_{i-1,j,k} + DT_{i-1,j+1,k} \\ & + ET_{i,j-1,k} + FT_{i,j,k} + ET_{i,j+1,k} + DT_{i+1,j-1,k} + ET_{i+1,j,k} \\ & + DT_{i+1,j+1,k} \quad (5) \end{aligned}$$

where

$$\begin{aligned} A &= \frac{1}{36} - \frac{1}{3}\theta & B &= \frac{1}{9} - \frac{1}{3}\theta & C &= \frac{4}{9} + \frac{8}{3}\theta \\ D &= \frac{1}{36} + \frac{1}{3}\theta & E &= \frac{1}{9} + \frac{1}{3}\theta & F &= \frac{4}{9} - \frac{8}{3}\theta \\ \theta &= \frac{K\Delta t}{2\rho C_p \Delta L^2} \quad (6) \end{aligned}$$

3. COMMENTS. Even though, the application of MWR-Collocation Yields finite-difference equations by the use of any of the following Laplacian term approximations ($\nabla^2 T$), this is not the case for FEDE:

$$L 16 = \frac{T_{i-1,j} + T_{i,j-1} + T_{i,j+1} + T_{i+1,j} - 4 T_{i,j}}{(\Delta L)^2}$$

$$L 17 = \frac{T_{i-1,j-1} + T_{i-1,j+1} + T_{i+1,j-1} + T_{i+1,j+1} - 4 T_{i,j}}{2 (\Delta L)^2} \quad (7)$$

$$L 19 = \frac{T_{i-1,j-1} + 4 T_{i-1,j} + T_{i-1,j+1} + 4 T_{i,j-1} - 20 T_{i,j} + 4 T_{i,j+1} + T_{i+1,j-1} + 4 T_{i+1,j} + T_{i+1,j+1}}{6 (\Delta L)^2}$$

$$L 17 19 = \frac{T_{i-1,j-1} + T_{i-1,j} + T_{i-1,j+1} + T_{i,j-1} - 8 T_{i,j} + T_{i,j+1} + T_{i+1,j-1} + T_{i+1,j} + T_{i+1,j+1}}{3 (\Delta L)^2} \quad (8)$$

If only L17 19 is chosen as the Laplacian approximation, one can prove that the MWR-Collocation yields finite difference equation and the MWR-Galerkin yields FEDE. The question now arises whether or not L17 19 is the best one because the unique FEDE can be obtained only by L 17 19 where as numerous MWR and FDM can be formulated. If accuracy, stability, and nonoscillation characteristics are derived, the order from best to least suitable is as follows:

Accuracy: L19 > L16 = L17 19 > L17

Stability: L17 > L1719 > L19 > L16

Nonoscillation: L17 = L1719 > L19 > L16

Certainly, the L17 19 is not the best one. Therefore, FEM may not be the best as far as accuracy, stability and nonoscillation characteristics are concerned. What happens if nonlinear variations are considered instead of linear nodal temperature variations within each element and also within each time increment? We do know nonlinear temperature distributions in space were considered in many finite-element codes. However, nonlinear variation with respect to time is not considered to the best of the author's knowledge. Has anyone derived FEDE with the assumption of nonlinear distribution in space? Do the conclusion with respect to accuracy, stability and nonoscillation characteristics remain valid for nonlinear variations? Further research is needed to answer some of these questions.

4. **STORAGE AND BANDWIDTH.** Everyone ends up with a system of linear algebraic equations irrespective of the use of any technique whether it is FDM, MWR or FEM. In the end, computers are used in association with some techniques for the solution of simultaneous linear equations. A smart way of resolving these equations is to use a banded matrix solution technique which has the advantages of speed and of using minimum computer core storage. This implies that one should ensure minimum matrix bandwidth. Indeed, one should also take advantage of sparseness (zeros) of a matrix. The following example illustrates the importance of minimization of core storage:

Table 1. Storage

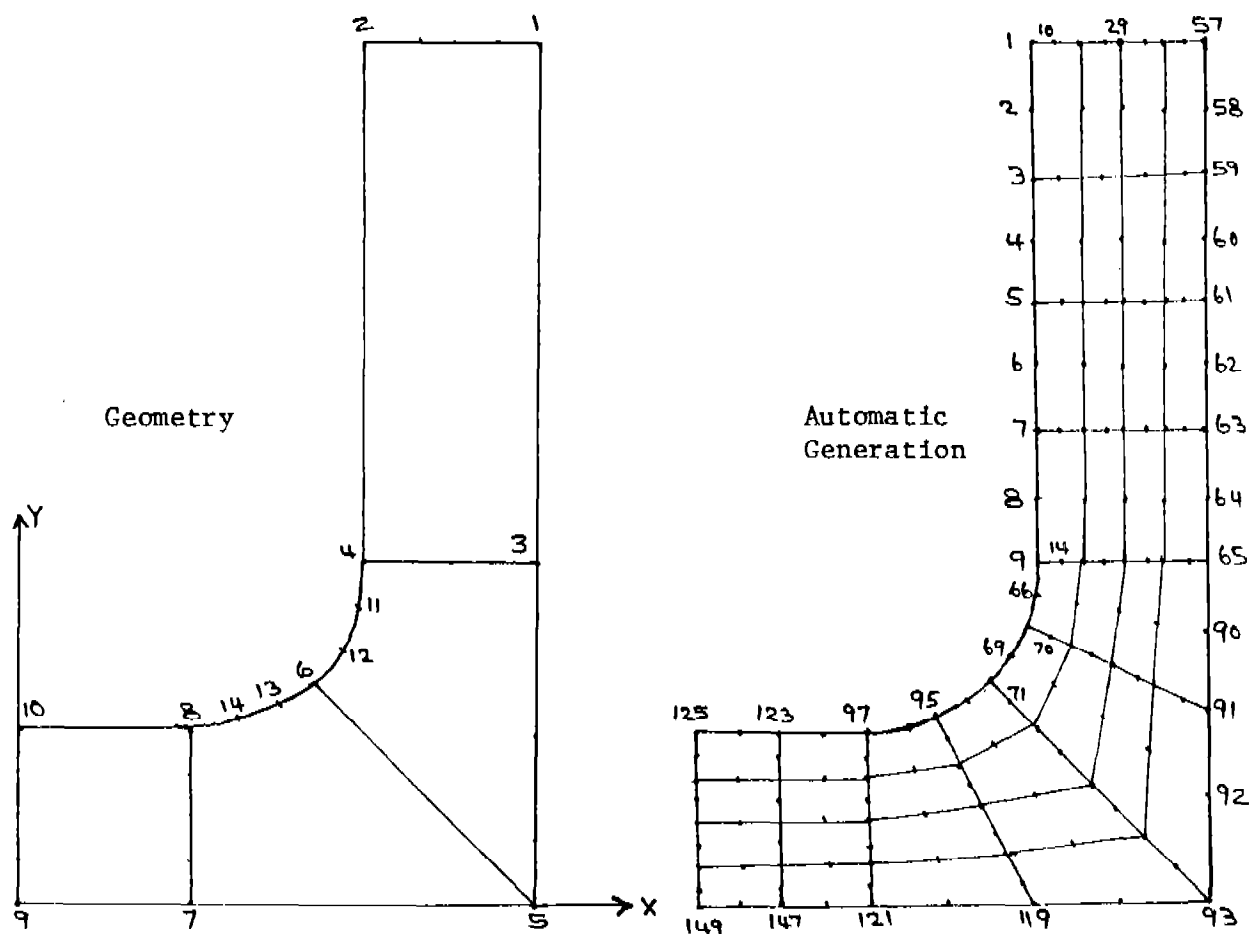
# of eqns	Full matrix	Upper Matrix	Upper Banded Matrix	Bandwidth
100	10,000	5,050	1,045	11
500	250,000	125,250	24,225	51
1000	1,000,000	500,500	49,725	51
			95,950	101

One may achieve the minimization of matrix bandwidth by numbering the nodal points in a structure in a particular manner. Either complex structure or inadvertent numbering yields an inefficient system of equations and thus one has to rely on automatic renumbering algorithms such as Cuthill-McKee. The following example illustrates its mechanics and effectiveness:

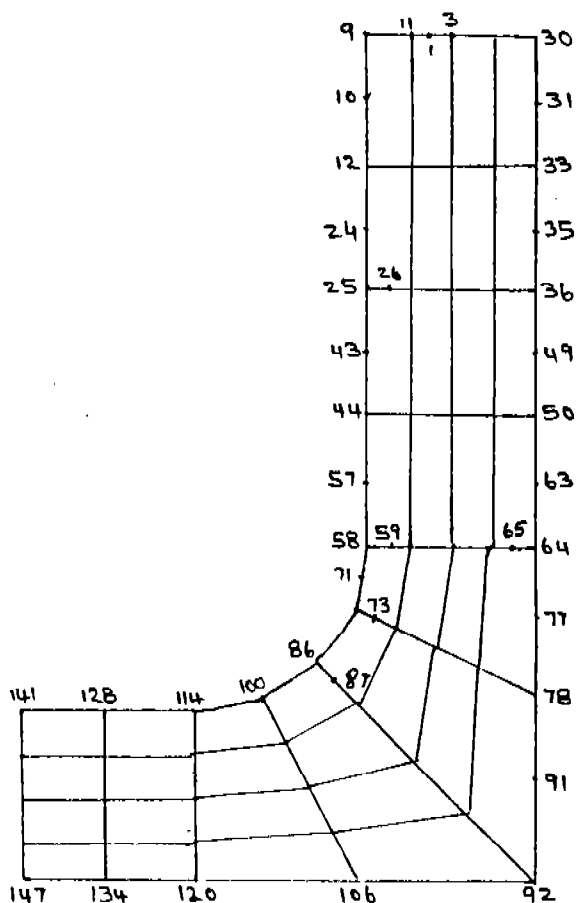
Table 2. Cuthill - McKee Technique

Configuration	Old Node #	Connected Node #	Degree	New Node #
<p>BEFORE</p> <p>Bandwidth = 6</p>	1	2,6,7	3	4
	2	1,3,7,8	4	7
	3	2,4,8,9	4	10
	4	5,3,10,9	4	13
	5	4,10	2	15
	6	11,1,12,7	4	2
	7	1,2,6,12,13,8	6	5
	8	2,3,13,14,7,9	6	8
	9	15,3,4,10,14,8	6	11
	10	5,15,4,9	4	14
	11	6,12	2	1
	12	11,6,7,13	4	3
	13	12,14,7,8	4	6
	14	15,13,8,9	4	9
	15	10,14,9	3	12
<p>AFTER</p> <p>Bandwidth = 4</p>				

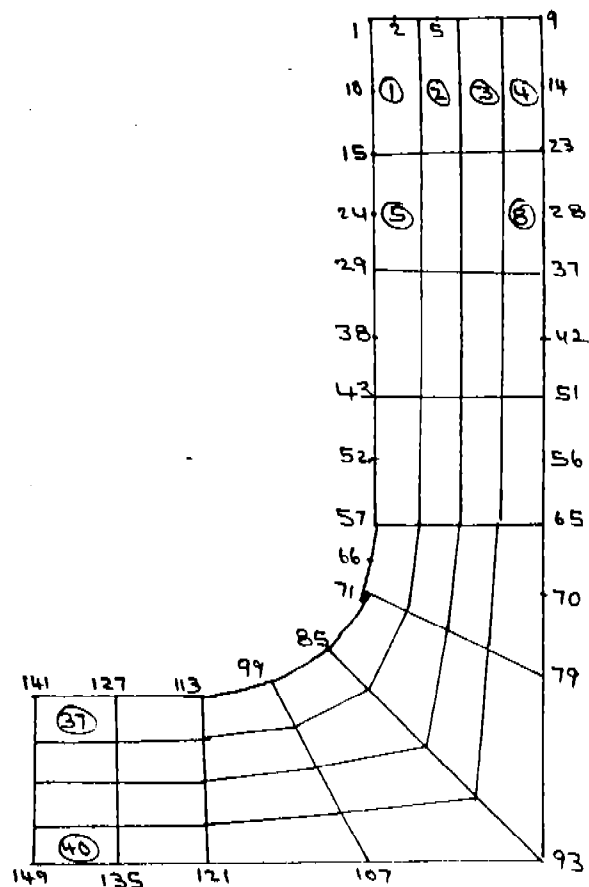
5. DISCUSSION. The importance of minimization of core storage and reduction of matrix bandwidth is demonstrated in Table 1. The automatic generation of mesh and automatic numbering and renumbering of nodes is quite common in most finite-element computer codes. The renumbering of nodes is accomplished, manually, by the Cuthill-McKee technique for a simple example. Fortunately, the minimum bandwidth is obtained in this case. Sometimes, this may not be true. Consider the following example with 40 quadrilateral (8-node) elements. Because of symmetry, this object may be considered as one-half of a crucible or rectangular channel filled partially by a hot molten metal. There are a total of 149 nodes. For example, the Marc Code generates the mesh and assigns the node numbers as shown for given geometry.



The automatically generated and numbered scheme does contain the bandwidth of 65. However, if an iterative (say 10 iterations) Cuthill-McKee algorithm is called for minimization of matrix bandwidth, the following renumbering is accomplished with a bandwidth of 26.



Cuthill-McKee



Manual

The bandwidth is 29, 29, 32, 35, 46, 29, 26, 26, 26, and 26 for iterations 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 respectively. Certainly, the final bandwidth is not an optimum choice. The manual numbering as shown above accomplished a bandwidth of 17 based on intuition. Thus, significant savings in both core storage space and computations is accomplished. Therefore, a prudent programmer/analyst, wherever possible, should not rely on automation alone.

6. REFERENCES.

1. Norrie, D., and de Vries, G., "Finite Element Bibliography," Plenum Press, NY 1976.
2. C. W. McCormick, "NASTRAN" Macneal-Schwendler Corp. Los Angeles, Calif.
3. Marc, MARC Analysis Research Corp. Palo Alto, Calif.
4. Bathe, K., "NONSAP" University of California, Berkeley, Calif.
5. Cuthill, E., and McKee, J., "Reducing the bandwidth of sparse Symmetric Matrices," pp.157-172, proc. of ACM National Conference, 1969.
6. Gurtin, M.E., "Variational Principles for Linear Initial Value Problems, "Quarterly J. of Applied Mathematics, Vol. 22, 1964, pp 252-256.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO Report 86-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Transactions on the Third Army Conference on Applied Mathematics and Computing		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Mathematics Steering Committee on behalf of the Chief of Research, Development and Acquisition		12. REPORT DATE February, 1986
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211		13. NUMBER OF PAGES 862
		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. The findings in this report are not to be construed as official Department of the Army position unless so designated by other authorized documents.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES This is a technical report resulting from the Third Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat various Army applied mathematical problems.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
robot vision	plate theory	
pattern recognition	conservation laws	
vulnerability	heat equations	
detection problems	stochastic problems	
control theory	finite difference methods	
mechanical systems	group methods	
stability problems	soliton equations	
flame theory	fusion processes	
bifurcation problems	viscoelasticity	
crack solutions	non-Newtonian flows	
viscoplasticity	Riemann problem	
Stefan problems	singular integrals	
fractal sets	elliptic systems	
penetration sets	tensor transformations	

